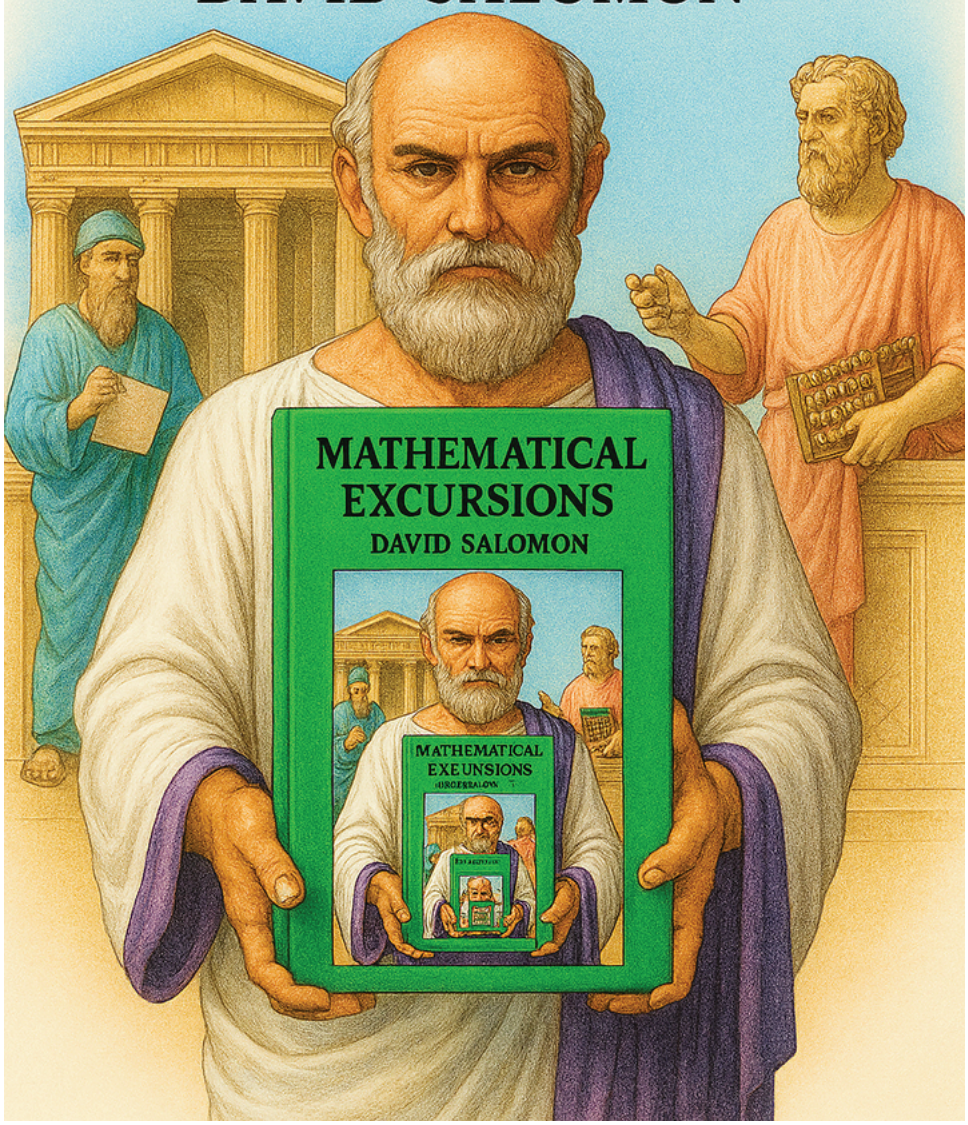
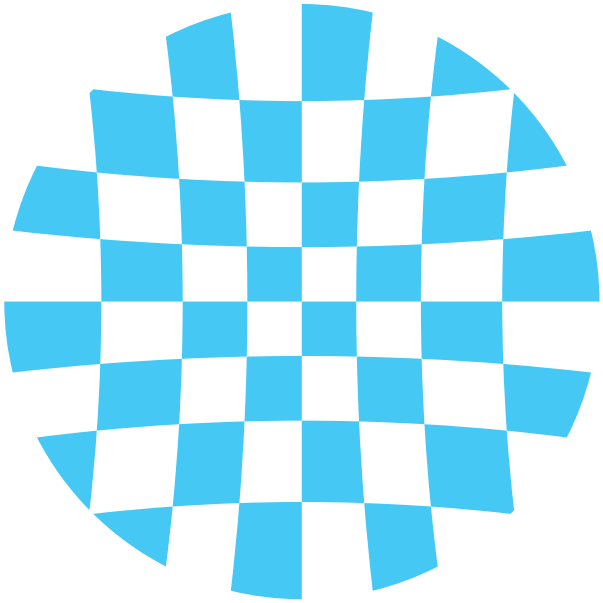
MATHEMATICAL EXCURSIONS

DAVID SALOMON



Dedicated to the seekers of beauty everywhere



I have long been thinking how to best illustrate the beauty of mathematics, and this book is the culmination of many attempts, thoughts, and trials that I have finally put on paper in an attempt to expose them to the silent scrutiny of cold print.

—The author.

The Beauty of Mathematics

In numbers' dance, a silent symphony plays,
A realm where logic and grace entwine their ways.
Infinite truths in finite forms concealed,
In mathematics, the universe revealed.

From zero's void to infinity's grand span,
The language of cosmos, known to man.
Prime numbers, the building blocks of all,
Stand alone, their secrets quietly call.

Patterns in nature, Fibonacci's code,
Golden ratios where beauty strode.
In fractals' maze, endless, self-repeating,
A glimpse of the divine in each small meeting.

Geometry's lines, precise and ever true,
Shapes and angles in harmonious view.
The circle, perfect, endless in its curve,
A symbol of unity, both calm and verve.

In algebra's realm, variables entwine,
Equations whisper, secrets to define.
Polynomials with roots complex and real,
Reveal a world beyond what senses feel.

Calculus, where motion meets still thought,
Rates of change in elegance are caught.
Derivatives, integrals, bound in their dance,
Describing worlds in a single glance.

Statistics, where chance and certainty blend,
Predictions made on data's trend.
The bell curve, its gentle rise and fall,
Mapping life's randomness, encompassing all.

In probability's paradoxical play,
Certainty in chaos finds its way.
Schrödinger's cat, both dead and alive,
Quantum realms where probabilities thrive.

Symmetry's allure, a mathematician's delight,
Mirroring nature in left and right.
Group theory's elegance, abstract and pure,
Invariants that through transformation endure.

Topology, where surfaces twist and turn,
Dimensions folded, beyond sight we learn.
The Möbius strip, with its single side,
Challenges the concept of space's divide.

Complex numbers, where real meets the dreamed,
Imaginary units, elegantly schemed.
Euler's identity, a marvel concise,
Where e, i, and π in unity splice.

Mathematics, a bridge 'tween earth and sky,
A quest for truth that never will die.
Each theorem a step on an endless ascent,
Towards understanding, our minds are sent.

The Pythagorean theorem, ancient and wise,
In right-angled triangles, its truth lies.
$a^2 + b^2 = c^2$'s might,
A beacon of logic in the night.

Fourier transforms, harmonies dissect,
Waves into frequencies, patterns we detect.
Signal processing, a world unseen,
Mathematics turns the abstract serene.

Game theory's strategies, choices arrayed,
In conflict and cooperation played.
Nash's equilibrium, a delicate balance,
Rational minds in competitive valance.

In cryptography, secrets safely hide,
Algorithms guard on data's tide.
Prime factors in RSA's firm hold,
Encrypting messages, stories untold.

Set theory, where infinity unfolds,
Cantor's diagonal, a paradox bold.
The continuum hypothesis, forever debated,
In the realm of the infinite, truth is fated.

Mathematics, an art, a science, a key,
Unlocking the secrets of reality.
From the simple to the complex intertwined,
The beauty of numbers, a gift to mankind.

In every equation, a story is told,
In every proof, a truth to behold.
Mathematics, a tapestry, woven with care,
A testament to the wonder we all share.

Thus, in the abstract, beauty does lie,
In the patterns and structures that defy
The limits of our sight, our mortal ken,
Mathematics, the purest of all human pen.

By `chatGPT 4o`, May 2024.

> Beauty is truth, truth beauty, that is all Ye know on earth, and all ye need to know.
> —John Keats.

Preface

Let's start with two imaginary (but very likely) conversations.

1. A typical conversation between two people on the street.

A. Who are you going to vote for?

B. Ms. so and so. She is first rate, as well as the best choice for our country. I'm certain of it.

A. What political party do you belong to?

B. The Nationalist party. We have the interest of the nation in mind. Yes, I know it for a fact.

A. How can we improve the economy?

B. That's obvious. The government has to increase taxes, subsidize the corn farmers, drop the interest rate by at least 2%, and eliminate the tariff on imported alcohol.

A. How much is 2% of 15 billion?

B. I don't know. This is math, and I've never been good at math. In fact, I've always hated it. Let's get back to the economy. I have more ideas.

2. Interviewing a candidate for the secret service.

A. You understand that this is a dangerous job? You might get killed.

B. I am ready to sacrifice my life for my country.

A. Good. Let's start. You are sent to the Gobi desert on a four-week secret mission. There are four agents under you. How much water should you take with you?

B. Just a minute. Let me get my calculator.

A. But this is only a simple multiplication!

B. This is math, and I've never been good with the multiplication table.

These types of responses are frequently seen. A lot of individuals hold firm beliefs on intricate matters like voting choices, preferred political parties, and economic enhancements. However, the same individuals are willing (even content) to acknowledge that they lack knowledge in mathematics and sometimes even ridicule or detest it. The quotes that follow are enlightening.

Man has always found it easier to sacrifice his life than to learn the multiplication table.

—W. Somerset Maugham, *Ashenden: Or the British Agent*, 1928.

This was something new and very interesting, and he felt that it was of greater import than trigonometry, which he never could understand.

—W. Somerset Maugham, *Of Human Bondage*, 1915.

I expect my mother's fear of decimal currency was related to her dislike of math, which is a common fear often dating back to a cruel teacher.

—Liane Moriarty, *Here One Moment*, 2024.

Pooh! Buck Mulligan said. We have grown out of Wilde and paradoxes. It's quite simple. He proves by algebra that Hamlet's grandson is Shakespeare's grandfather and that he himself is the ghost of his own father.

—James Joyce, *Ulysses*, 1922.

Every morning I thank heaven for giving me another day without algebra.

—Unknown.

Three out of two people have trouble with fractions.

—Unknown.

When people are asked why they hate math, the most common answer is "Math isn't always taught well." The following answers are only slightly less common: (1) math is hard, even impossible, (2) it is only for bright kids, (3) no fun at all, (4) it is logical, answers are either right or wrong, (5) it involves numbers like 1, 2, 3 and so on; confusing, (6) it's unnecessarily formal, (7) because math teachers are failed engineers, (8) I am not hardwired for math. (9) I hate it simply because it is MATH!



Figure 1: Image by chatGPT 4o, Aug. 2025.

Parents, teachers, school principals, and mathematicians have long been aware of the disturbing fact that many people claim to hate ALL of mathematics, whereas it is rare to hear someone claim to hate all of literature or all of art. There must be something in mathematics that alienates certain types of people. However, the modern world is based on science and engineering, disciplines where mathematics plays an essential role, which is why the question of what to do about math anxiety is so important. It is my hope that this book will prove a small step toward a solution. The book tries to alleviate math anxiety by illustrating those aspects of mathematics that can be considered unexpected, powerful, or beautiful. It also tries to present mathematics as an integral part of the overall scientific endeavor, and to give the reader a glimpse of the nature of mathematical thinking.

Today, the number of professional mathematicians in the world (in schools, universities, industry, research institutes, and in government service) runs into the hundreds of thousands. There are also many math enthusiasts who spend their time and are having fun in studying mathematics, solving problems, and trying to understand difficult concepts. Someone who hates math would simply say that these people are weird, but the math professionals and amateurs claim that they love their field because it is beautiful, interesting, and full of surprises and unexpected results. In general, it is fun. Far from being a dry and esoteric subject, they feel that mathematics is a rich and living part of our culture and that the essence of mathematics lies in its elegance and the intellectual challenge it offers.

> Work is much more fun than fun.
> —Noel Coward.

Math is a human language, much like the many other languages that are written and spoken by humans. It is a language because it makes it possible for people to communicate. This has always been true, even in the distant past, because even then people needed to communicate in order to interact, trade, cooperate, and even to fight. Monetary transactions, land surveying, construction of buildings and monuments, and organizing armies, are examples of tasks that require communications and cooperation.

Mathematics is also a sense, a sense that allows us to perceive things in the real world, just like vision and touch make it possible for us to sense parts of the world, so math is a sense for patterns, relationships, and logical connections between facts and statements. As with other senses, people have different amounts of each, which is perhaps why so many have a weak sense of math. If you need eyeglasses, should you claim I am not hardwired for seeing? And yet many people, young and old claims just that when it comes to math, and thereby deprive themselves from sensing a large part of the world.

Why do so many people hate mathematics? I would like to propose the following reasons:

Reason 1. Mathematics is hard. A person coming across a mathematics text with an interesting title may open it, browse a few pages, examine the many bizarre symbols and unfamiliar expressions, emit a deep sigh, and put the book away as too difficult. In contrast, a mathematician who finds himself in the same situation may often say to himself, this is certainly difficult, but I consider it a challenge and I'd like to spend the time and effort to master it and eventually obtain much enjoyment from it.

Reason 2. The tremendous progress that was made in this field in the last two or three centuries. This progress led to abstraction, to a way of thinking with logic but without pictures, that is foreign to most people.

Reason 3. Mathematics is seen as the science of numbers and numbers are dry, boring, and only good for calculations. Mathematicians, however, know that their discipline is about much more than just numbers and computations. Mathematics is not numerology, it is not the domain of accountants, and it consists of much more than arithmetic. Today, mathematics is seen as the study and understanding of *patterns*. Here is a short historical survey of the development of this field of knowledge.

I have always found mathematics to be challenging and unengaging. The abstract nature of mathematical concepts, coupled with the repetitive and rigid processes required to solve problems, does not resonate with my strengths and interests. I prefer tasks that allow for creativity, flexibility, and a focus on language or visual elements, rather than the structured and formulaic approach that mathematics demands. This has led to a lack of enthusiasm and confidence in my mathematical abilities, ultimately making the subject unenjoyable for me.

—A short explanation of why I hate maths, by `chatGPT 4o`, June 2024.

I love mathematics because of its logical structure and the clarity it brings to complex problems. The ability to break down intricate issues into solvable steps through well-defined rules and formulas is both satisfying and intellectually stimulating. Mathematics provides a universal language that helps explain patterns and relationships in the world around us. I enjoy the challenge of solving mathematical problems and the sense of accomplishment that comes with finding precise solutions. This discipline enhances my problem-solving skills and analytical thinking, making it a deeply rewarding subject for me.

—A short explanation of why I love maths, by `chatGPT 4o`, June 2024.

The term "Mathematics" is derived from the Greek. It starts with $\mu\alpha\nu\theta\acute{\alpha}\nu\omega$ (mantáno, "I learn"), to which is appended the suffix $\mu\alpha$ (ma), which changes its meaning to "that which is learned."

It is true that mathematics started as the study of numbers. Up to about 500 B.C., Egyptian, Babylonians, and other mathematicians were concerned with solving practical problems that involved numbers (mostly integers). Greek mathematicians, generally active in the period from 500 B.C. to A.D. 300, were mostly interested in geometry and how to measure the areas and volumes of geometrical figures. Thus, Greek mathematics was concerned with both numbers and shapes. In the middle ages, mathematicians came to terms with the concepts of irrational numbers, complex numbers, and the modern (Hindu-Arabic) notation still in use today. They were also interested in solving equations of high degrees, but it was not until modern times—specifically, the 1600s—that new concepts were included in mathematics. The ideas and contributions of Descartes, Newton, Leibniz and others had extended mathematics by introducing the concepts of space, motion, and rate of change.

Also in the 1600s, Blaise Pascal made important contributions to probability, the study of chances and their behavior, thereby introducing a new type of pattern to mathematics, the pattern of chance. In the early 1800s, the young Evariste Galois introduced the concept of groups, thereby adding symmetries and structure to the patterns of mathematics. In the late 1800s, Georg Cantor introduced the transfinite numbers.

Since the late 1800s, many types of mathematical objects and patterns—such as tensors, vector spaces, point sets, reasoning (logic), and fractals—have been introduced and became legitimate objects of study and research.

> A mathematician, like a painter or a poet, is a maker of patterns. If his patterns are more permanent than theirs, it is because they are made with ideas.
>
> —G. H. Hardy, *A Mathematician's Apology*, 1992.

The word "pattern" is a fortunate choice in this context because it connotes order, regularity, and lawfulness and because a pattern is abstract. The pattern of floor tiles, for example, is not the tiles themselves but an abstract attribute associated with the tiles and with the way they are laid on the floor.

Thus, modern mathematics is a diverse discipline and it can be considered the study of many different types of patterns, chief among them are patterns of counting, shape, motion, change, symmetry, structure, position in space, regularity, reasoning, chance, and infinity. In addition, mathematics helps us understand the world we live in. For a not-yet-fully-understood reason, the behavior of physical objects can concisely and accurately be described, and even predicted, by mathematical expressions. Thus, thanks to mathematics, we are better able to understand the universe, the objects in it and how they interact, and to communicate this understanding to others in ways that reveal the unity and order that we discover in complex phenomena.

The fact that mathematics is extremely useful for solving and explaining real-world problems has always been and still remains a puzzle. The short article [Wigner 60] is just one of many publications where this unexpected effectiveness of mathematics is discussed.

It is important to say a few words about mathematical existence. Do numbers really exist? What does it mean to say that a certain mathematical object, relation, or theorem exists? Today, mathematicians agree that mathematical existence means only logical self-consistency. For a mathematical object to exist, it does not have to have physical reality. If we make a list of non-contradictory axioms and rules, and then use it to deduce results, then these results exist mathematically. Thus, existence in mathematics is similar to positions in chess. They follow from previous positions and rules, but do not have physical existence the way everyday objects have.

> The great book of nature can be read only by those who know the language in which it was written. And this language is mathematics.
>
> —Galileo Galilei.

This is certainly not the only book that tries to encourage an interest in mathematics. There are many texts that discuss the magic, mystery, beauty, and fun of math with various degrees of success. Some of these texts try to accomplish this task without the use of equations, formulas, expressions, and symbols. The idea being that those who need help with mathematics would be discouraged by formulas and may not even try

to benefit from the book or at least to enjoy it. However, this book is not shy about using and displaying many mathematical objects, because I believe that notation is important and that mathematics is easier to master and enjoy when the proper notation is understood and used. Anyone who doubts this claim should try to add and multiply two 2-digit integers expressed in Roman numerals.

(Modern mathematical notation makes it easier to write, understand, and manipulate expressions, but beginners often find it daunting. The notation is compressed and also requires a strict syntax. A short mathematical expression of just a few symbols may contain a great deal of information about variables, constants, and operations.)

> Someone told me that each equation I included in the book would halve the sales. I therefore resolved not to have any equations at all. In the end, however, I did put in one equation, Einstein's famous equation $E = mc^2$.
>
> —Stephen Hawking, *A Brief History of Time: From the Big Bang to Black Holes.*
>
> Equations are more important to me, because politics is for the present, but an equation is something for eternity.
>
> —Albert Einstein.

◇ **Exercise 1:** A book with 20 equations has sold just one copy. How many copies would it have sold if it had no equations at all?

When and how will I use the math taught at school?

We use math all the time, but we often do so subconsciously, without realizing what it is that we actually do. At school, a child may struggle with the multiplication table for a few years, but eventually he gets used to doing simple multiplications in his head, forgetting that what he is doing quickly and automatically is the same math that caused him so much anguish in the past. When the child grows up, he may believe that he never uses math and that math is useless, while all the time he is using elementary arithmetic, which is part of math.

Those who graduated from elementary school may not realize this, but each time they come across quantitative information—a graph, a numeric expression, percentages, probabilities, or statistical facts in the news—they use the mathematical background acquired during their school years to understand, make sense of, and analyze this quantitative (and therefore mathematical) information. We can claim that all the difficult and boring math lectures in class, as well as the math homework after school, have developed in our mind the mental ability to deal with quantitative information without realizing that what we are doing is an embodiment of math, the math that we believe we dislike.

This is why I say that one answer to the common question "when will I use all this math?" is "you will use it all the time, without realizing that you use it."

Here are several examples of using elementary mathematical operations in real life:

1. Shopping in a store, you discover ketchup on sale. This is an item that your children like, so you always keep enough at home. You stop and consider how much to buy. This depends on the number of children, how much they use, how much space there is in your pantry, and how deep the discount is. A matter of estimates and addition/multiplication.

2. You decide to paint several rooms in your house. The amount of paint needed depends on the total area of the walls and on the paint coverage; how much area will one gallon of paint cover. Again, this requires only a few operations, but it is a common, practical application of math.

3. A child gets permission and a budget to allocate part of the backyard for his own private garden. He needs to decide what to grow and how to fit his preferences to his budget. This may require division in addition to multiplication.

> I knew I should have become a plumber.
> —A frustrated math teacher.

It is obvious that when such a math problem is part of a child's math homework, it seems boring and impractical, but years later, when the child faces a similar problem as an adult, his background springs into action and helps him solve the problem without realizing its connection to the old, despised homework. Many teachers tend to assign such practical math problem as homework, not realizing that (1) these problems tend to move the students away from math in the short run, while (2) will help them in the future.

The problem of convincing students that math can and will be useful to them in the future becomes more acute in the higher grades, when teaching algebra, logarithms, and especially calculus, a topic that deals with rates of change. For these topics, it is difficult to find examples and problems drawn from everyday life. Most applications of solving equations and computing derivatives have to do with scientific and engineering applications, and they leave the average student thinking "if this is how algebra is used for, then it has nothing to do with me. I will never use it." Examples and problems of this kind only serve to drive the average student away from mathematics.

Thus, when a student asks in an algebra lesson "how is this useful?" A teacher can often answer only "because you will need it in next week's exam." An answer that proves to the student even more the impracticability of mathematics.

Often, we realize that we don't know certain things. We admit that we don't know how to fix our car, how to fill out the confusing tax forms, how to cook oxtail, and why we have to reset our clocks twice a year or why some years are leap years. On the other hand, there are many things, facts, and topics that we don't know and we don't realize that we don't know.

In the grate the flickering embers
Served to show how dull November's
Fogs had stamped my torpid members
Like a plucked and skinny goose.
And as I prepared for bed, I
Asked myself with voice unsteady,
If of all the stuff I read, I
Ever made the slightest use.

James Clerk Maxwell "A Vision of a Wrangler, of a University, of Pedantry, and of Philosophy" in [Cambell and Garnett 82], p. 307

Sometimes the answer to ignorance of mathematics is "but you can easily look it up." The point is that you can look up only what you realize that you don't know. If you don't know that a topic exists, you cannot look up anything in it. This is a good reason why we should study mathematical fields that we may never use.

A student doesn't know what topics he does not know, which is why he may conclude that he already knows everything that is important and applicable to his future career. As a result, the student may choose classes that he likes, and forgo other topics, not knowing and not realizing that those topics exist and may later prove useful to him in his future work. This is another reason why topics such as algebra and calculus should be taken by students and be treated seriously.

> Every thought that one permits through his mind leaves its trace. Thoughts are things. Our lives are governed a great deal by our thoughts.
>
> —Spencer W. Kimball in [Kimball 06].

Before we look at any details of mathematical beauty and elegance, here is an example that beautifully illustrates the topic of this book. Most mathematicians agree that the well-known Euler equation

$$e^{i\pi} + 1 = 0$$

is beautiful, and it is easy to see why. It connects the five most-important numbers in mathematics 0, 1, e, π, and i in a compact expression. It reveals an unexpected connection between them, and it also includes the important operations of addition, multiplication, and exponentiation.

> The laws of nature are written in the language of mathematics.
>
> —Galileo.

The Unreasonable Effectiveness of Mathematics in the Natural Sciences

As a young student in elementary school I vividly remember my surprise when, perhaps in 3rd grade, the science teacher wrote the first equation on the blackboard. I don't remember what elementary concept he was trying to explain, perhaps something to do with pulleys or with speed and motion. He suddenly wrote an equation that

related very nicely to the topic under discussion, and followed it with more equations and expressions (a mathematical equation is in expression that contains an equal sign and an unknown).

I was startled, as if something unexpected has happened; not necessarily wrong, but surprising. Why are we allowed to use mathematics to explain the world? What is the justification for this? Why does the world we happen to live in follow mathematical logic? Or is it the opposite, is our mathematics an inevitable result of the particular world we live in? Heady questions! Even at such a young age I felt that this practice needed to be justified.

Over time, after seeing and using many equations, I got used to this way of doing and describing science, and then I discovered that I was not alone. Other people have also been asking the same question, namely why is mathematics so effective and successful in describing the behavior of the universe. The discussion that follows is a summary of explanations of and solutions to this question that others have been proposing.

The title above is due to a much discussed, intriguing 1960 article by Eugene Wigner, reference [Wigner 60]. In it, Wigner discusses and analyzes the question of why the real world, the world we see around us, can be so effectively described by mathematical expressions, equations, and functions. The root of the problem is that mathematics is a creation of the human mind, whereas the world is something we are born into. We try to unravel and understand its behavior and the laws it obeys, but we have no control over it. Reference [WikiWigner 23] lists several solutions proposed by scientists and philosophers, and here is a summary of the relevant ideas.

Note. None of the ideas and solutions discussed here is ultimate or complete, but each may prove satisfactory to some readers. The question treated here may be one that does not have a full answer, at least not at our present stage of knowledge.

We start with a summary of Wigner's article. Many mathematical ideas were originally developed in contexts far removed from science, but are routinely used today to illustrate modern scientific concepts and theories. Why is this justified? Perhaps the chief example is the ellipse, which was discovered by the ancients while considering the conic sections, and then used by Newton to mathematically describe the orbits of planets.

Sometimes, a scientist uses very little experimental evidence to claim a new theory or discovery, and then writes a mathematical function or relation to formally "explain" it. Based on his work, the discovery is tested again and again until, after passing many tests, it is accepted as a theory.

An excellent example is the 1929 discovery of the expansion of the universe by Edwin Hubble. Hubble's paper that announced the astonishing fact contains his famous diagram where only 6–7 of the 21–22 galaxies shown are on or very close to the straight line that proves the expansion. Nevertheless, it is this unconvincing diagram which has quickly convinced the scientific community that we live in an expanding universe.

Maxwell's equations are an excellent example of important scientific predictions driven by mathematics. It was Maxwell who noticed that the mysterious constant c in his equations has dimensions of speed. He therefore conjectured c to be the speed of propagation of both the electric and magnetic fields. The actual electromagnetic waves were discovered in 1879 and used by Hertz in 1886 to send information wirelessly. Special relativity, published in 1905, employs the fact that c is a constant to claim that

the speed of light is constant. It is a pity that Maxwell himself did not live to see those fundamental implications of his equations.

Since 1960, several people have responded to Wigner's article, trying to solve and explain his basic question. In 1975, Hilary Putnam explored several arguments, among them the following: Perhaps the use of mathematics to formally describe physical phenomena is not mysterious because it is simply wrong. After all, in our world, it is humans who check and interpret the results of other humans, and being human, they cannot be objective. Thus, scientific results and their mathematical interpretations are not objective. The mathematics is wrong, and its use is therefore not mysterious.

In 1980, Richard Hamming came up with four arguments that even he admitted were unsatisfactory.

1. Science is based on experiments, not, as the ancient philosophers believed, simply on thinking. A theory, even the simplest and most basic, is wrong if it is disproved by even a single experiment. However, humans often see what they are looking for.

A famous example is the canals on Mars. In 1877 astronomer Giovanni Schiaparelli observed the planet Mars and was surprised to "see" what he later described as canali (channels). As so often happens, something was lost in the English translation and canali came out as canals. The belief in Martian canals was strong and persisted until 1964, when NASA's Mariner IV photographed the planet and found no canals. I distinctly remember reading about the canals in science books for children in the mid 1940's.

Thus, a long lasting scientific theory was based on what astronomers wanted to see. Any mathematics based on the canals—such as how long they took to dig, how many Martians worked on them, and how much water they transferred—must also have been false; they explained nothing.

2. Science can perhaps be explained by mathematics, but the world that we experience consists of much more than science, and those parts of the world cannot be explained by mathematics.

3. Scientists choose what mathematical tools they need to explain a given theory. They also often extend mathematical concepts and theories as needed to explain new physical discoveries, concepts, and theories. We can therefore claim that mathematics is a tool that grows with our growing knowledge of the world. In some sense, mathematics complements science, which is why the two are so compatible.

4. We humans find it natural to think of facts mathematically, because of the way in which natural selection, working blindly, has developed us from our ancestors. Therefore, it is natural for us to use mathematics to explain what we observe in nature.

Michael Atiyah has proposed to turn the tables and look at this mystery as the unreasonable effectiveness of physics in mathematics. Perhaps mathematics, the mental creation of humans, is the dominant fact in the universe, and we should ask why the world conforms to our mental creation in such a way that scientific theories can explain the mathematical theories that we have come up with, sometimes already in ancient times.

Perhaps the simplest explanation so far, due to Max Tegmark, is that the physical world is completely mathematical. This assumption gets rid of the mystery in a most natural way.

> to properly know mathematics is to know the only thing that can be assuredly known. Politics and sociology and psychology have facts you have to interpret, but mathematics facts are just facts. There is no arguing. There is no left-wing or right-wing algebra. There is no sin in geometry, and no guilt in trigonometry.
>
> —Matt Haig, *The Life Impossible*, 2024.

Is mathematics a science?

The connection between mathematics and science is an intriguing question that has been a source of debates, arguments, misunderstandings, and disagreements for centuries. The explanations and examples shown here are only meant to illuminate the concepts of what is science and what is mathematics. They are not going to settle this question. This section is based on reference [arachnoid 08].

We start with commonly accepted definitions (from Wikipedia):

"Mathematics is a field of study that discovers and organizes methods, theories and theorems that are developed and proved for the needs of empirical sciences and *mathematics* itself."

This definition has two shortcomings: (1) It considers mathematics a field of discovery, whereas some may consider mathematics a process of invention. (2) This definition defined mathematics partly in terms of itself (in italics).

"Mathematics involves the description and manipulation of abstract objects that consist of either abstractions from nature or, in modern mathematics, purely abstract entities that are stipulated to have certain properties, called axioms."

This seems a better definition, mostly because it mentions the connection between mathematics and nature. We often naively assume that mathematics is a creation of the human mind and has nothing to do with the world we perceive around us, but here are two examples that illustrate connections between the two.

1. The famous spirals found in sunflowers and their relation to the Fibonacci sequence. See Page 295.

2. Cicadas in their prime, or the connection between cicadas and prime numbers. Certain species of periodical cicadas have life cycles that last exactly 13 or 17 years, which are both prime numbers.

The Wikipedia definition of science. "Science is a systematic discipline that builds and organizes knowledge in the form of testable hypotheses and predictions about the universe."

These definitions are from Wikipedia. I believe that science can better be defined by the following statements:

1. Science is a field of knowledge that collects data from experiments and observations.

2. It then attempts to develop theories based on this data and tries to extend these theories and generalize cases.

3. New data can demolish (falsify) any theory. Any attempt to believe in a theory against new contradictory data is unscientific.

Notice that statements 1 and 2 can be swapped. It is possible, even common, to first create a theory and then try to find support for it in nature. A good example of this approach to science is general relativity, which was published by Einstein in 1916 without any experimental evidence to support it. On the other hand, the theory of

natural selection was published by Charles Darwin in 1859, many years after collecting the relevant data in the Galapagos islands during 1831–1836. For more than 20 years Darwin was thinking, developing ideas, writing, and incubating his data before finally making it public.

Statement 3 is most important, because it distinguishes science from superstition. This statement is the essence of the important *falsifiability criterion* which has long been recognized as an essential constituent of science. A scientific theory can never be proven correct, but can easily be proven wrong by even a single experiment that contradicts it. A superstition, on the other hand, cannot be disproved, because it is based on belief, not on evidence.

> Science in itself is a sweet beginning and sometimes a bitter ending, whereas superstitions have neither beginning nor end; it is complete in itself.
>
> —From *Times of India*, March 2020.

Notice that scientific data must be unambiguous, verifiable (other labs and researchers should be able to repeat it), and objective (anyone who sees it should come to the same conclusion). However, the precise way data is obtained is unimportant. Different researchers may collect the same data from different sources. Radioactivity, for example, was discovered (by Henri Becquerel in 1896) by experimenting with Uranium, but those who came after Becquerel discovered the same natural phenomenon by working with Thorium.

Once these definitions and statements are understood, we can discuss the differences between, and the similarities of science and mathematics.

Science must be falsifiable. A scientific theory can be proved wrong, but cannot be proved correct. On the other hand, a mathematical statement can be proved wrong (by a counter-example), but can also be proved correct. The concept of a proof, in fact, is a central pillar in mathematics. We can prove by logic that a claim, a statement, or a theorem, is true, and this fact constitutes a big difference between mathematics and science. It suggests that the two are different disciplines in principle.

As an example, there are countless (well, hundreds of) proofs of the Pythagorean theorem about the sides of a right-angle triangle. Mathematicians, engineers, and students feel strongly that no one would ever be able to refute this theorem. It is eternal and does not depend on any future scientific discoveries or progress.

On the other hand, mathematical statements and conjectures can be demolished by experiment, although fast computers are often needed for this. A simple example is the prime generator function $x^2 - x + 41$ discovered (produced? guessed?) by Euler in 1772. Today, with the help of computers, it is easy to show that this function has a very limited range. It produces prime numbers for integer x values from 0 to 39, but fails at $x = 40$. Hence, a simple computation proves that this isn't a universal prime generator. In this respect, mathematics and science seem similar.

Another example of a mathematical claim that was difficult to prove wrong by logic but was actually easy to falsify by computations, is the well-known Monty Hall problem (Page 341), which is based on a subtle probabilistic argument. The naive solution that was proposed by thousands of viewers and readers proved false by an expert, but could have been shown to be false if those believers had written a simple computer program to simulate the process. The result would have shown immediately that their proposed

naive solution was wrong.

Would you consider a logician a scientist? There are logical statements that can be, and that have been, falsified. On the other hand, logicians are viewed by the rest of us as the purest of pure mathematicians, people who do not perform computations, do not write software, rarely deal with numbers, and never consider numbers that are non-integers. And yet logic is a bona-fide field of mathematics. In this respect, mathematics and science are viewed as different disciplines.

Must science be based in nature? Science fiction is a literary genre whose practitioners claim: "Science fiction is not about future technology and light sabers. It is about life in a different society or even a different universe. In principle, it is possible to develop a science based on an imaginary universe where the laws of nature are different from ours. If those laws are consistent and free of contradictions, a science based on such a world would be as good as our science. It may not be the "real" science, but it may be as rich, complex, and logical as our science, and may also be falsifiable. In this respect, mathematics may be such a science. We may consider it a system of imaginary objects, their properties and relations between them, and such a system may not be very different from an imaginary science based on a fictional world.

> Math could be surprisingly emotional because most of the time it wasn't.
> —Lee Child, *Eleven Numbers: A Short Story*, 2025.

More about mathematical proofs. In the past, proofs were produced mostly by professional mathematicians and sometimes by amateurs. In the 20th century, the world was several times surprised by proofs that were either obtained with the help of computers or are so long and convoluted that only a few experts can understand and appreciate them. This suggests that as the world of mathematics is progressing and becomes more complex and abstract, we are likely to see more examples of such proofs, a trend that makes some uneasy. Here are two well-known examples of such proofs:

1. The four color map theorem states that a maximum of four colors is required to color the regions of any map so that no two adjacent regions have the same color. Figure 2 illustrates what is meant by adjacent regions. Regions that share a single common point, as in part (a) of the figure, are not considered adjacent, but if their common boundary is longer, as in part (b), they are adjacent and require different colors.

It seems that it was Augustus de Morgan who first became interested in this problem and made it popular, but it wasn't until 1976 that it was solved, by Kenneth Appel and Wolfgang Haken, with the aid of elaborate computer software. This problem is considered historically important because it was the first major theorem to be proved using a computer. Needless to say, their proof was not immediately accepted by the mathematical community. Some experts felt that mathematical proofs should be produced by brain power alone, while others pointed out that such a vast piece of software (it had to individually check a set of 1,476 possible map configurations) was infeasible for a human to check manually. Today, the consensus is that the proof is correct.

2. From Wikipedia: Fermat's Last Theorem states that no three positive integers a, b, and c satisfy the equation $a^n + b^n = c^n$ for any integer value of n greater than 2. The cases $n = 1$ and $n = 2$ have been known since antiquity to have infinitely many solutions.

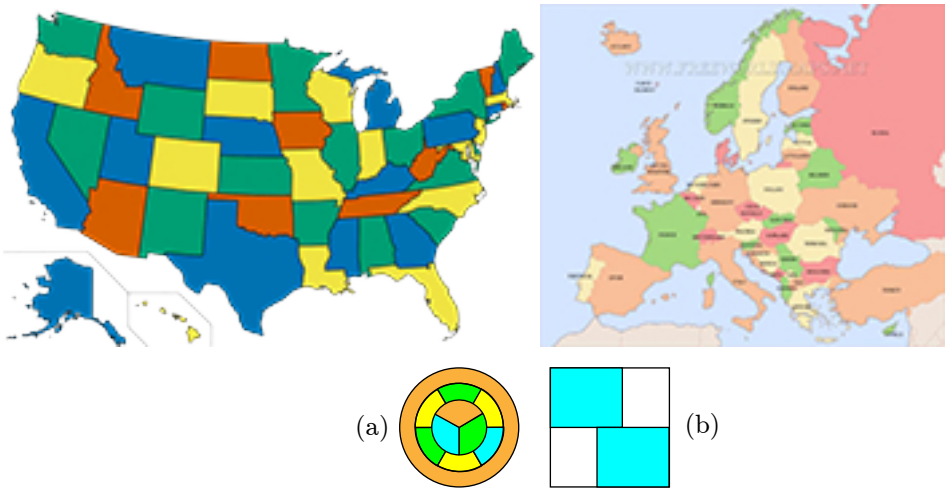


Figure 2: Map Coloring.
USA map licensed by "creativecommons.org/licenses/by-sa/3.0/deed.en."

◇ **Exercise 2:** The ancient Greeks already knew that the square root of 2 is irrational. Use Fermat's Last Theorem to prove that the cube root of 2 is also irrational.

Pierre de Fermat, a French lawyer and judge, was also a serious amateur mathematician. Today, he is considered one of the two leading mathematicians of the first half of the 17th century. (The other is René Descartes.) It was around 1637 that Fermat wrote the now-famous note in the margin of his copy of *Arithmetica* (an ancient text on number theory by Diophantus). "It is impossible to separate a cube into two cubes, or a fourth power into two fourth powers, or in general, any power higher than the second into two like powers. I have discovered a truly marvelous proof of this, which this margin is too narrow to contain."

Those who followed Fermat took this note seriously, because Fermat was known to make correct statements and conjectures while omitting their proofs. However, 358 years of efforts by many mathematicians yielded no proof, until Andrew Wiles published a 129-page-long proof in 1993. After a few corrections in 1994, this long and esoteric communication has come to be regarded as a formal proof of the theorem.

> Fermat "was a mathematician of rare power. He was an independent inventor of analytic geometry, he contributed to the early development of calculus, he did research on the weight of the earth, and he worked on light refraction and optics. In the course of what turned out to be an extended correspondence with Blaise Pascal, he made a significant contribution to the theory of probability. But Fermat's crowning achievement was in the theory of numbers."
>
> —Peter L. Bernstein, *Against the Gods*, 1996.

The history of mathematics has many examples of long and complex proofs that were later simplified, often several times, by their originator as well as others. It would therefore come as mo surprise if Wiles's proof would be greatly simplified in the future,

but when it first came out, several mathematicians objected to it in principle, claiming that such long proofs can be checked and understood only by few specialists, and should therefore remain suspect until they are simplified and shortened.

The discussion and examples in this section will not put an end to the debate about whether mathematics is a science. They can only serve as an aid and guidelines to those interested in this question.

About This Book

This book was started as a collection of beautiful facts, objects, theorems, and relations in mathematics. Over time, however, as more and more material was added, it became simply a place to summarize, discuss, and explain mathematical topics that are of personal interest to me. As a result, the book is personal (some may say that it is a hodgepodge of topics and facts). An occasional reader may find certain topics that are of interest and may skip the rest of the book. In any case, it is free.

This book contains many examples, exercises (all with answers), references, and jokes (see index). The references have the format [name date] and are listed in detail in the bibliography at the end of the book. In 2024, in response to the growing "intelligence" of chatGPT and its cousins, I now include poems, images, and texts generated by various AI sources. These addition are only meant to amuse and entertain the reader.

> Not to put ourselves in a still more false position, we had better to explain things with an example, and perhaps the benevolent reader will understand precisely what our difficulty is, the more so as this example will not be a digression, but, on the contrary, a direct and immediate continuation of the story.
>
> —Fyodor Dostoevsky, *The Idiot*, 1869.

Any errors, mistakes, misprints, and mistypes found here, as well as any criticism and suggestions, can be emailed to me at `dsalomon@csun.edu`.

The single most compelling reason to explore
the world of mathematics is that it is beautiful.
—Calvin C. Clawson, *Mathematical Mysteries*.



Mathematical Jokes

A joke is a display of humor in which words are used within a specific and well-defined narrative structure to make people laugh and is usually not meant to be interpreted literally.

A mathematical joke is a form of humor which relies on aspects of mathematics or a stereotype of mathematicians. The humor may come from a pun, or from a double meaning of a mathematical term, or from a lay person's misunderstanding of a mathematical concept.

—From Wikipedia

Math is like love; a simple idea, but it can get complicated.

Dear algebra, please stop asking us to find your X. She's never coming back, and don't ask Y.

Hey, have you ever noticed what's odd? Every other number!

All non-parallel lines in a plane meet once and never again.

I was perpendicular to a girl I liked a lot. Never saw her again after high school.

You have dialed an imaginary number, please rotate your handset 90° and try again.

Parallel lines never meet, but have much in common.

Draw them on a spherical surface for a happy ending.

Our friend the ray spends its entire life traveling with no real destination.

Non-straight lines must be gay.

Owlgebra is the type of math birds like best.

π says to i, "get real." i says to π, "be rational."

The minus sign: "Are you sure I make a difference?" The other sign: "I'm positive."

Which tool is best for math? The multi-pliers.

There are only 10 types of people, those who understand binary, and those who don't.

Quadratics hit rock bottom, then find their way to greatness.

Negative quadratics spend their life to get to the top, only to then drop all the way.

For us, x is unknown, but for the Romans `X` was always 10.

Why was six afraid of seven? Because seven, eight, nine.

Sine waves also have their ups and downs.

The moon and six pence is a misnomer. Rather, the moon has four quarters.

How does it feel to be sixteen? "It feels really square, but next year I'll be in my prime."

How do you turn seven into an even number? Remove the "s."

How can you argue with a 90° angle? It is always right.

Daddy, why is the math book so sad? Because it has so many problems.

Circles are pointless.

A physicist, a biologist and a mathematician are sitting in a cafe, watching people entering and leaving a house across the street. First they see two people entering the house. Then they notice three people leaving the house. The physicist says, "The measurement wasn't accurate." The biologist says, "They must have reproduced." The mathematician says, "If one more person enters the house, it will be empty."

Mathematical Quotations

A quotation is the repetition of a sentence, phrase, or passage from speech or text that someone has said or written.

—From Wikipedia

Pure mathematics is, in its way, the poetry of logical ideas.

Without mathematics, there's nothing you can do. Everything around you is mathematics. Everything around you is numbers.

Mathematics is the most beautiful and most powerful creation of the human spirit.

For many people, math is a four-letter word.

The most painful thing about mathematics is how far away you are from being able to use it after you have learned it.

We will always have STEM with us. Some things will drop out of the public eye and go away, but there will always be science, engineering, and technology. And there will always, always be mathematics.

Mathematics as an expression of the human mind reflects the active will, the contemplative reason, and the desire for aesthetic perfection. Its basic elements are logic and intuition, analysis and construction, generality and individuality.

As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

What is mathematics? It is only a systematic effort of solving puzzles posed by nature.

Mathematics is the music of reason.

Mathematics knows no races or geographic boundaries; for mathematics, the cultural world is one country.

There should be no such thing as boring mathematics.

God used beautiful mathematics in creating the world.

This result is too beautiful to be false; it is more important to have beauty in one's equations than to have them fit experiment.

"Obvious" is the most dangerous word in mathematics.

Mathematics are the result of mysterious powers which no one understands, and which the unconscious recognition of beauty must play an important part. Out of an infinity of designs a mathematician chooses one pattern for beauty's sake and pulls it down to earth.

Mathematics allows for no hypocrisy and no vagueness.

I've always enjoyed mathematics. It is the most precise and concise way of expressing an idea.

It is impossible to be a mathematician without being a poet in soul.

A mathematician who is not also something of a poet will never be a complete mathematician.

Mathematics compares the most diverse phenomena and discovers the secret analogies that unite them.

Mathematics is not about numbers, equations, computations, or algorithms: it is about understanding.

Somehow it's okay for people to chuckle about not being good at math. Yet, if I said "I never learned to read," they'd say I was an illiterate dolt.

In mathematics the art of proposing a question must be held of higher value than solving it.

Beauty is the first test; there is no permanent place in the world for ugly mathematics.

It is clear that the chief end of mathematical study must be to make the students think.

Go down deep enough into anything and you will find mathematics.

Nature is written in mathematical language.

Mathematics has beauty and romance. It's not a boring place to be, the mathematical world. It's an extraordinary place; it's worth spending time there.

To me, mathematics, computer science, and the arts are insanely related. They're all creative expressions.

Why do children dread mathematics? Because of the wrong approach. Because it is looked at as a subject.

Many who have had an opportunity of knowing any more about mathematics confuse it with arithmetic, and consider it an arid science. In reality, however, it is a science which requires a great amount of imagination.

The study of mathematics, like the Nile, begins in minuteness but ends in magnificence.

Wherever there is number, there is beauty.

Life is a math equation. In order to gain the most, you have to know how to convert negatives into positives.

Mathematics may not teach us to add love or subtract hate, but it gives us hope that every problem has a solution.

One of the endlessly alluring aspects of mathematics is that its thorniest paradoxes have a way of blooming into beautiful theories.

Mathematics is not a careful march down a well-cleared highway, but a journey into a strange wilderness, where the explorers often get lost. Rigor should be a signal to the historians that the maps have been made, and the real explorers have gone elsewhere.

The pure mathematician, like the musician, is a free creator of his world of ordered beauty.

Just because we can't find a solution, it doesn't mean there isn't one.

Mathematics is a place where you can do things which you can't do in the real world.

Millions saw the apple fall, but Newton asked why.

The definition of a good mathematical problem is the mathematics it generates rather than the problem itself.

If I were again beginning my studies, I would follow the advice of Plato and start with mathematics.
Pure mathematicians just love to try unsolved problems, they love a challenge.
I've always been interested in using mathematics to make the world work better.
I hope that seeing the excitement of solving this problem will make mathematicians realize that there are lots and lots of other problems in mathematics which are going to be just as challenging in the future.
The only way to learn mathematics is to do mathematics.
It's fine to work on any problem, so long as it generates interesting mathematics along the way, even if you don't solve it at the end of the day.
You don't have to be a mathematician to have a feel for numbers.
Sometimes the questions are complicated and the answers are simple.
The essence of math is not to make simple things complicated, but to make complicated things simple.
If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.
Mathematics is a game played according to certain simple rules with meaningless marks on paper.
Dear Math, please grow up and solve your own problems. I'm tired of solving them for you.
I am still waiting for the day I'll use mathematics integration in real life.
Arithmetic is numbers you squeeze from your head to your hand to your pencil to your paper till you get the answer.
That awkward moment when you finish a math problem and your answer isn't even one of the choices.
Mathematics is like love; a simple idea, but it can get complicated.
If there is a 50-50 chance that something can go wrong, then nine times out of 10 it will.
Math is fun. It teaches you life and death information like when you're cold, you should go to a corner since it's 90 degrees there.
Pure mathematics is the world's best game. It is more absorbing than chess, more of a gamble than poker, and lasts longer than Monopoly. It's free. It can be played anywhere, Archimedes did it in a bathtub.
Mathematics consists of proving the most obvious thing in the least obvious way.
In mathematics, you don't understand things. You just get used to them.
There are two ways to do great mathematics. The first is to be smarter than everybody else. The second way is to be stupider than everybody else, but persistent.
Five out of four people have trouble with fractions.
Mathematics is a hard thing to love. It has the unfortunate habit, like a rude dog, of turning its most unfavorable side towards you when you first make contact with it.

What's your favorite number?

This question may have as many answers as there are responders, and they (the answers) vary widely from jokes, to nonsense, to intriguing, all the way to useful. Here are a few such answers, randomly found on the Internet.

- 987654321/123456789 = 8.0000000729. Useful when you need a rational approximation of 8.
- "nan" (not a number), because I get it often in my calculator.
- 1729, the famous Ramanujan's taxicab number.
- 91, the smallest number that is a sum of two perfect cubes in two different ways $3^3 + 4^3$ and $-5^3 + 6^3$. Notice that 91 also equals $9^0 + 9^1 + 9^2$ and is also the sum of Fibonacci numbers $F_1^1 + F_2^2 + F_3^3 + F_4^4$. There is also the claim (unverified) that $2^{91} = 2475880078570760549798248448$ is the largest power of 2 that does not contain the digit 1. Some believe that 91 looks like a prime and smells like a prime, but is not a prime.
- 24, 36, and 60, in this order, because they have many divisors.
- $355/113 \approx 3.14159292035$, because it provides a useful approximation of π.
- "Umpteen." It sounds big, yet it is in the teens.
- To me, 2 is THE number, because computers use binary numbers.
- 808,017,424,794,512,875,886,459,904,961,710,757,005,754,368,000,000,000, the order of the monster group.
- 867-5309, known as Jenny's number.
- To me, the favorite number has always been ϵ (epsilon).
- 2520, because it is the first natural number that is divisible by the decimal digits 1 through 10.
- 0 and 1, two favorites.
- 17, because of the wallpaper symmetry groups and also the surprising compass-and-straightedge construction of the regular 17-gon discovered by Gauss. (17 is also a factor of 100,000,001.)
- 6, because it is the first perfect number, as well as common in nature, as described in Exercise 57.
- 137, because 1/137 is the approximate value of the fine-structure constant, a measure of the strength of the electromagnetic force.
- 10, because it is the base of our number system.
- I like to show and tell that 433222111 equals one four, two threes, three twos, and one one. (See also Pages 69 and 303.)
- ϕ, the golden ration. Need I say more?

- 6741. Kaprekar's constant. It seems magical.
- 625. Why? Are you asking why? Because it is the largest fourth power of a prime that doesn't end in 1.
- 13. Don't know why I like it.
- 911 is in my opinion the most important number is the USA.
- I have always been fascinated by -1, the first negative integer.
- $\sqrt{2}/2 = 1/\sqrt{2} \approx 0.7971$, is the first of the many constants that show up often in electrical calculations, as well as in trigonometry.
- Graham's number. One of the biggest numbers ever used in a mathematical problem. This number is so incredibly large that it is practically impossible to express it in standard notation. Even the number of zeros it requires is beyond comprehension, even for mathematicians. Using Knuth's up-arrow notation, Graham's number, the brain child of Ronald Graham, is g_{64}.
- Zero, because it is both a number and nothing.
- 113383, because it is the smallest number whose Collatz expansion overflows java's double precision.
- 42. This number is the answer to the Ultimate Question of Life, the Universe, and Everything.
- 13, because it is associated with Friday, my favorite day.
- Obviously, π should be everyone's favorite.
- 2, because it is the first and only even prime.
- Sextillion, 10^{21}, because it rolls off the tongue so smoothly.
- $256 = 2^8$. It is so common when working with computers and software.
- 2^{32} was also commonly used to express capacities of drives and RAM.
- 69. It has such a rare rotational symmetry.
- One more than the largest number mentioned earlier.

Definitions of mathematics:

Contemporary style:
Mathematics is the study of patterns, structures, quantities, and relationships using abstract reasoning and logical analysis.

Style of Shakespeare:
Mathematics is the noble art where reason doth entwine with number, shape, and harmony to unveil the hidden truths of nature's grand design.

Johnny Carson's style:
Mathematics, folks, is that amazing thing where you take numbers, twist them around, and somehow convince yourself that it all makes sense, kind of like my tax return!

—By ChatGPT-4o, May 2025.

Contents

Preface ix

Introduction 1

1 Graphics: Visible Math Objects 19

1.1 Curves and Surfaces 20
1.2 Perspective 38
1.3 Moiré Patterns 39
1.4 Ruled Surfaces 41
1.5 Most Important Curve 41
1.6 Listings of Mathematica codes 42

2 Numbers: The Basic Building Blocks 47

2.1 Arithmetic Operations 47
2.2 Logical Operations 49
2.3 Integers 50
2.4 Rationals and Irrationals 78
2.5 Continued Fractions 85
2.6 Real Numbers 98
2.7 Complex Numbers 103
2.8 Hypercomplex Numbers? 112
2.9 Quaternions 115
2.10 Transcendental Numbers 129
2.11 Important and Interesting Numbers 130
2.12 Complex Golden Ratios 156
2.13 Approximating Formulas 163
2.14 Cyclic Numbers and Metadromes 164

3 Symmetry **167**

3.1 A bit of History 169
3.2 Symmetry Groups 170
3.3 Orbifold Notation 189
3.4 The Magic Theorem 196
3.5 Orbifold Examples 197
3.6 Two-Dimensional Transformations 198
3.7 Symmetry in Tiling 222
3.8 Tessellations 228
3.9 Circle Inversions 230
3.10 Symmetry in text, speech, and ... 232

4 Infinity **239**

4.1 A Short History of Infinity 240
4.2 Mathematical Infinity 241
4.3 Potential and Completed Infinities 243
4.4 Unexpected Results of Infinity 247
4.5 Set Theory 251
4.6 Physical Infinity 259
4.7 Infinitesimals and the Calculus 261
4.8 Restrictions in Mathematics 267
4.9 Debates in Mathematics 272

5 Order: Sequences and Series **275**

5.1 Equations 276
5.2 The Pythagorean Theorem 277
5.3 A Different Dirac Equation 279
5.4 Sequences 280
5.5 Numerical Sequences 281
5.6 The Fibonacci Sequence 285
5.7 Metallic Ratios 300
5.8 The Comma Sequence 301
5.9 Quasi-Numeric Sequences 303
5.10 Series 304
5.11 The Real Harmonic Series 309
5.12 The Book-Stacking Problem 312

6 Paradoxes **317**

6.1 Types of Paradoxes 317
6.2 Examples of Paradoxes 319

7 Probabilities: the Rule of Chance 345

7.1 Basic Concepts 345
7.2 More Probability Concepts 349
7.3 Randomness 352
7.4 Benford's Law 361
7.5 Randomness in Dice 376
7.6 Go-First Dice 380
7.7 Subjective Probability 382
7.8 Probability and Psychology 383
7.9 The Birthday Paradox 385
7.10 Choosing a Candidate 387
7.11 Examples of Unexpected Probabilities 389
7.12 Probabilistic Counting and HLL 402

8 Geometry 409

8.1 Fractals 410
8.2 Weierstrass Function 427
8.3 Continuity 429
8.4 Interpolation 434
8.5 Least Squares Interpolation 435
8.6 Perlin Noise 440
8.7 Three-Dimensional Rotations 450
8.8 Vector Operations 451
8.9 Points and vectors 469
8.10 Representing Curves 472
8.11 PC Curves 474
8.12 Polynomial Interpolation 477
8.13 Spline Interpolation 481
8.14 Hermite Interpolation 483
8.15 Interactive Control 484
8.16 The Hermite Curve Segment 485
8.17 The Cubic Spline Curve 490
8.18 Cardinal Splines 495
8.19 Parabolic Blending: Catmull-Rom Curves 499
8.20 Bézier Approximation 503
8.21 The Bézier Curve 503
8.22 The Bernstein Form of the Bézier Curve 506
8.23 Linear Perspective 510
8.24 Perspective: Basic Concepts 530
8.25 The Mathematics of Perspective 537
8.26 Slanted Squares with Integer Corners 547
8.27 Area of regular polygons 548
8.28 The Fourth Side of a Triangle? 549

9 Puzzles 551

9.1 Examples of Puzzles 551

10 Miscellaneous topics 581

10.1 The Gamma Function 581
10.2 Magic Squares 583
10.3 Parking as a greedy problem 589
10.4 Error-Control Codes 591
10.5 Compact Disc (CD) 596
10.6 Reed–Solomon Codes 598
10.7 SVD Image Compression 604
10.8 What is Average? 607
10.9 The power of the XOR 608
10.10 Brouwer fixed-point theorem 611
10.11 Short Topics 613

Bibliography 621

Answers to Exercises 633

Index 692

The content of most textbooks is perishable, but the tools of self-directness serve one well over time.

—Albert Bandura.



Introduction

The following discussion illustrates how a bright child, someone who has just been introduced to integers and basic arithmetic operations, may discover a beautiful mathematical relation. The child knows that adding certain pairs of integers, such as 1 and 9, 2 and 8, and 5 and 5, produce a sum of 10. It may occur to him to multiply the two integers of each pair, in order to find out which pair yields the largest product. The results are summarized in Table 1a. It is clear that the pair $(5, 5)$ yields the largest product, 25, but it is also easy to see that the other four pairs yield products (in column 3) that become smaller in a regular way. Column 4 lists the differences between 25 and the products in column 3, and column 5 lists the same differences as squares of consecutive integers. Not an earth shattering discovery, but a discovery nonetheless.

1	2	3	4	5
5	5	25	0	0^2
4	6	24	1	1^2
3	7	21	4	2^2
2	8	16	9	3^2
1	9	9	16	4^2

(a)

1	2	3	4	5
6	6	36	0	0^2
5	7	35	1	1^2
4	8	32	4	2^2
3	9	27	9	3^2
2	10	20	16	4^2

(b)

Table 1: Adding and Multiplying Five Pairs.

Most children may not be interested in such a game. Some children may stop when they make this simple discovery, but others may ask themselves, is this pattern specific to 10? It is easy to show (Table 1b) that the integer 12 features the same pattern, and after spending a few years at school, a bright child can use algebra to prove, as shown here, that this pattern of behavior is general

$$n + n = 2n, \quad n \times n = n^2. \quad (n - i) + (n + i) = 2n, \quad (n - i)(n + i) = n^2 - i^2.$$

Such an achievement is very satisfying and may result in the child becoming, if not a professional mathematician, at least someone with an understanding of and an interest

in mathematics.

Mathematics is absolute, infinite, and eternal. A mathematical result that has been satisfactorily proved, whether positive or negative, is certain; there is no doubt about it. Mathematics is the only area where we come face to face with absolute certainty as well as with the infinite. Everything else that we experience in this world is finite.

> Two things are infinite: the universe and human stupidity; and I'm not sure about the universe.
>
> —Albert Einstein.

Mathematics is timeless, as every mathematical fact, outcome, or theorem will endure indefinitely. The results achieved and confirmed by ancient civilizations remain accurate in society today, just as they did upon their initial discovery. Furthermore, they will remain accurate regardless of wars, revolutions, beliefs, advancements in science, or any upcoming social or historical occurrences. For this reason, it can be argued that mathematics is the sole endeavor through which one can achieve eternal fame. An idea, outcome, or principle in mathematics that was originally introduced or proved by someone like Professor Henry Higgins cannot be invalidated or demonstrated by another individual. It remains true and it secures Higgins's position in history. Before we explore further, here are various illustrations of the concept of beauty in mathematics along with unique mathematical characteristics and properties not seen in other fields.

> A thing of beauty is a joy for ever:
> Its loveliness increases; it will never
> Pass into nothingness; but still will keep ....
>
> —John Keats, *Endymion*, 1818.

■ **Boring numbers**. We start with the concept of a boring number. Some may claim that every number is boring, but consider the following. The integer zero means nil, zilch, nothing, naught, and no thing. Yet it is not boring and is even important. The concept of zero, acting as a place holder, made it possible to replace the cumbersome system of Roman numerals with the much superior Hindu-Arabic numerals that we use today. The integer one is the first positive integer and is therefore of interest. The integer two is the first even integer, and so is not boring. Three is the first odd integer. Four has the interesting property that it equals both the sum and the product of two twos, and so is not boring. Section 2.3 lists many examples of interesting numbers.

This leads us to the little-known number paradox [WikiInteresting 23], a humorous but logical argument. Imagine that we partition the integers into two classes, interesting $\mathcal{I}$ and boring $\mathcal{B}$, according to certain criteria. The set $\mathcal{B}$ of boring integers may be infinite, but it must have a smallest member $\mathcal{S}$. $\mathcal{S}$ is the smallest boring number, but this fact makes it interesting, so we have to move it to set $\mathcal{I}$. After the move, if set $\mathcal{B}$ is non-empty, it must have another smallest number, which should also be moved to $\mathcal{I}$ for the same reason. When this process is completed, we find that set $\mathcal{B}$ is empty; there are no boring numbers.

■ **Real numbers**. We are familiar with the integers. They are easy even for little children to visualize and comprehend. When a child asks How many numbers are there, the answer, Infinitely many, may come as a surprise, but it is easy to accept when the

Of Boring Numbers

In chamber dim, where candle wan doth flicker,
A scholar sits, his brow with burden bowed.
The text pale, with ciphered rows grows thicker,
Yet none among them sings nor speaks aloud.

What joy hath "seven" in its crooked stance?
What tale doth "four" in silent corners keep?
They march in line, devoid of song or dance,
A lifeless throng that lulls the soul to sleep.

"Lo!" quoth he, "the sum of sums is sorrow,
No jest, no jest in figures grim and cold.
They reckon not the dreams of tomorrow,
Nor whisper truths the ancient bards foretold."

The plus and minus, root and power squared,
Are chains that bind the mind in rigid frame.
No muse e'er kissed a digit, nor declared
Her love for logic's ever-burning flame.

Yet still he writes, as duty doth compel,
Though heart be numb and ink be void of fire.
For in the count, some hidden truth may dwell–
A spark to lift the soul from mire to lyre.

So let the numbers drone their dull refrain,
Whilst poets plot escape through verse again.
Yet in the hush, where digits dare not speak,
The soul may stir, and hidden meaning seek.



—Poem and image by MS Copilot, Sept. 2025.

child hears the following argument. It is possible to add 1 to any number x and in this way, end up with a bigger number. Thus, there is no largest number and there are infinitely many numbers. However, when the real numbers are introduced at school a few years later, they are much harder to accept, to deal with, and to understand. The real numbers include the integers, the rationals (numbers of the form m/n where both m and n are integers) and the irrationals. (The latter are numbers with infinitely many digits after the decimal point, where the digits follow no pattern. Most of the roots and logarithms belong in this set.)

One conceptual difficulty with the real numbers is that they are dense. We know that there are infinitely many integers, but they are discrete. Each integer has a successor and a predecessor, and there are gaps between them that are full of real numbers. Thus, there are infinitely many real numbers between the integers 1 and 2, but there are also infinitely many real numbers between, say 1.6 and 1.7, as well as between 1.6 and 1.60001.

We therefore say that the reals are dense in any interval. Between any real numbers a and b there are infinitely many real numbers.

A surprising result of this property of the reals is that a real number does not have an immediate successor or a predecessor. Given a real number a, there is no number that immediately follows or precedes a. If we find a number b that looks like the immediate successor of a, we may eventually realize that the number c, defined as $(a + b)/2$ is located midway between a and b. Similarly, the number $d = (a+c)/2$ is located midway between a and c, and so on.

An unexpected corollary is that there is no such thing as the smallest positive real number. Such a number, the immediate successor of zero, does not exist; it forms a hole in the real number line. Another unexpected result (a paradox?) is that there are infinitely many such holes in the number line, but at the same time the line must also be continuous. On one hand we visualize this line as a set of discrete points, each corresponding to a real number but, because the reals are dense, we cannot imagine any gaps between the points on this line, nor can we imagine any missing points. We therefore refer to the real number line as the continuum, and we simply live with the fact that the real numbers raise conceptual problems that we may never solve or understand.

A set of mathematical objects known as surreal numbers was introduced into mathematics in 1969 by John H. Conway. Among other features, surreal numbers attempt to extend the concept of real numbers. Thus, every real number is surrounded by surreals, which are closer to it than any real number. Also, between any two real numbers there are infinitely many surreals. Do the surreal numbers exist? A mathematical object is a matter of definition. Anyone can come up with a set of such objects and they would become part of the mathematical universe if they are consistent and free of contradictions, and especially if they extend a previous mathematical idea, theory, or object. The definition of surreal numbers consists of a notation (how to write them) and rules of arithmetic operations on them. For detailed information, see the wikipedia article `Surreal_number`.

- The field of modern analytic geometry was independently developed by Pierre de Fermat and René Descartes around 1637. This field is an important illustration of the beauty of mathematics, because it shows how to translate between numbers and geometric figures and objects. Instead of trying to describe the shape of the top of Mount Everest, we can simply tell the reader to locate the point with coordinates $27^\circ\,59'\,1''$ N $86^\circ\,55'\,31''$ E on his map.

- Figure 10.16, copied here, illustrates another aspect of the beauty of mathematics. The corners of a cube can be numbered such that each digit position in a corner number corresponds to a different dimension of the cube. We can consider the square a two-dimensional cube, because it is one face of a three-dimensional cube. Similarly, a straight segment can be viewed as a one-dimensional cube, constructed as a face of a two-dimensional cube.

  The figure shows how the corners of cubes of any dimensions can be numbered in this way, and makes it clear how to number, for example, the 16 corners of a four-dimensional cube, even though we cannot visualize such an object. We have therefore translated cubes (part of geometry) into numbers, which makes it easy to imagine and analyze higher-dimensional objects without having to deal with pictures.

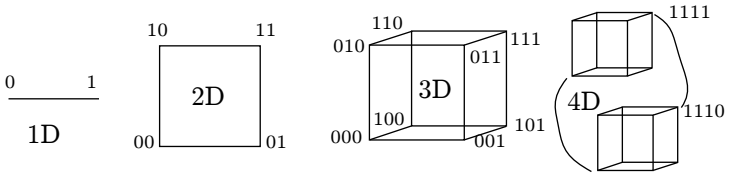


Figure 10.16. Cubes of Various Dimensions and Corner Numbering.

■ Here are two proofs, geometric and algebraic, that $\sqrt{2}$ is irrational. The first one is described in [Apostol 2000]. See also Page 82.

Figure 2 shows an isosceles right triangle ABO. We assume that its two equal sides AB and BO are of length 1 and its hypothenuse AO, whose length, according to Pythagoras, is $\sqrt{2}$, is a rational number of the form m/n, where m and n have no common divisors (i.e., they are coprime or relatively prime). We multiply each triangle side by n to end up with a hypothenuse of length m. Now all three sides are integers and they are the smallest set of integers for this triangle, making it the smallest possible isosceles right triangle whose sides are all integers. If we try to shrink a side by any factor, some sides would become non integers.

We now construct the circular arc from B to C to obtain AB = AC. Since both AO and AC are now integers, CO must also be an integer. We draw segment CD as the perpendicular to AO at point C to obtain point D, and observe that CDO is also an isosceles right triangle and therefore CD = CO, implying that triangles ABO and CDO are similar. Because hypothenuse AO is rational, hypothenuse DO must also be rational, but this contradicts the assumption that m and n are relatively prime (in lowest terms).

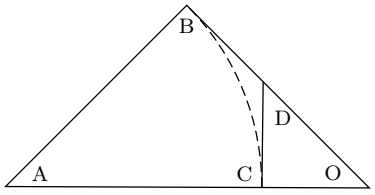


Figure 2: Try a Rational $\sqrt{2}$.

Many texts list the ancient proof, provided by the Greeks, which is arithmetical, so it is a surprise to see that there are also other approaches to prove the same fact, and they are based on different mathematical concepts and and techniques.

> "How many proofs of the Pythagorean Theorem do you know?" Erdős asked. "One," Vázsonyi said. "I know thirty-seven."
>
> —Andrew Vázsonyi in Paul Hoffman, *The Man Who Loved Only Numbers*, 1998.

As a bonus, here is another, short algebraic proof. Assume that $\sqrt{2} = p/q$ where p and q are positive integers and q is the smallest possible. This implies that $p - q > 0$

while $p - 2q < 0$. We can now write

$$\frac{2q-p}{p-q} = \frac{2-p/q}{p/q-1} = \frac{2-\sqrt{2}}{\sqrt{2}-1} = (2-\sqrt{2})(\sqrt{2}+1) = \sqrt{2},$$

which contradicts the assumption that q is the smallest possible.

> Math was abstract and eternal, unchanging, discovered not invented, but when human beings used it, they always left a fingerprint.
>
> —Lee Child, *Eleven Numbers: A Short Story*, 2025.

■ **Infinity in math**. Infinity plays an important part in many mathematical areas, where it can be extremely useful. The human mind, however, cannot grasp infinity, so we have to be wary when infinity is used to solve a problem, prove a theorem, or support a claim or an idea. Here are examples of the use of infinity in mathematics, and how wrong results are easy to obtain when this important concept is used carelessly.

1. The sum of the infinite series $C = 1 - 1 + 1 - 1 + 1 - 1 + \cdots$ can be computed in the following ways. First, we can write it as the sum of differences $C = (1 - 1) + (1 - 1) + (1 - 1) + \cdots$, where it adds up to zero. Then we can write it as the sum $C = 1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \cdots$, where its sum is 1, and finally as $C = 1 + (-1 + 1 - 1 + 1 - 1 + \cdots) = 1 + (-C)$, which implies $2C = 1$ or $C = 1/2$. These three sums are different and all employ the fact that C is infinite. Which is correct? It is obvious that the partial sums obtained when we add more and more terms of C oscillate between zero and 1 and would never deviate from these values. This justifies the conclusion that series C has no unique sum or, more accurately, does not converge to a sum.

Note. Given a series $\sum_{i=1}^{\infty} a_i$, it is always possible to construct the sequence whose kth element is the partial sum $\sum_{i=1}^{k} a_i$. It is known that if this sequence converges, then its associated series also converges, and vice versa. Applying this concept to series C above, we construct its partial-sums sequence 1, 0, 1, 0, 1,... which does not converge. Being both tricky and devious, mathematicians may point out that the average of this sequence converges to 1/2. Thus, even though series C diverges, we can somehow consider the value 1/2 as a representative of its sum.

2. Consider the infinite fraction $D = 0.999\ldots$, which we denote by $0.9\overline{9}$. The quantity D cannot be greater than 1, so it must either be equal to or less than 1. Even though it contains infinitely many digits, it is easy to show that D equals 1, and here are three proofs (see also Page 100). The first proof says that if D were less than 1, there would be a real number $F = (D + 1)/2$ between it and 1, but F cannot exist because D contains infinitely many digits and because all those digits are 9's, the largest decimal digit. This kind of proof is often referred to as hand waving and may not be rigorous.

The second proof start with $D = 0.999\ldots$, continues with $10D = 9.999\ldots$, and subtracts $9D = 10D - D = 9.999 - 0.999 = 9$, from which it is obvious that $D = 1$. A straightforward use of the infinite nature of D.

The third proof is geometric and is based on the area of a circle. Figure 3 shows a circle of area a. We cut a hole at the center, whose area is $0.1a$. The remaining ring has an area of $0.9a$. We now repeat this process. We think of the first hole as a new circle of area $0.1a$ and cut a smaller hole, of area $0.01a$ at its center, leaving a small ring

of area $0.1 - 0.01 = 0.09a$. When this is repeated without limit, the total areas of the rings, which is the infinite sum $0.9a + 0.09a + 0.009a + \cdots = 0.999a\ldots$, must equal the original area a, which proves that $0.9\overline{9} = 1$.

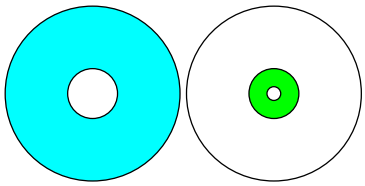


Figure 3: Circles and Rings.

This fact is one of many surprises and unexpected results that appear in mathematics. Several such surprises appear in this book, such as the Khinchin constant (Page 62).

3. We now apply the technique of the previous proof to the "opposite" of D, the huge number $G = \ldots 9999.0$. We first compute $10G = \ldots 99990.0$, and then add 9 to both sides, obtaining

$$10G + 9 = \ldots 99990.0 + 9 = \ldots 99999.0 = G,$$

or $10G + 9 = G$, a simple equation whose solution is $G = -1$. This clearly makes no sense and is the result of treating the infinite carelessly when we multiplied both sides by 10 in step 1, simply by shifting G to the left. We conclude that the use of infinity in mathematics can be beneficial, but results obtained with the help of this quantity should be verified before they can be considered valid.

4. Finally, a proof that avoids the naive approach of treating infinity as a number. Instead, it uses the epsilon-delta (ϵ, δ) technique of approaching a limit. We start by constructing the infinite sequence

$$x_1 = .9,\ x_2 = .99,\ x_3 = .999,\ \ldots,\ x_n = \sum_{k=1}^{n} 9 \cdot 10^{-k} = 0.\underbrace{99\ldots9}_{n},$$

and define D as the limit of this sequence, *if it exists*.

In order to prove that the limit is 1, we examine the difference

$$|x_n - 1| = |0.\underbrace{99\ldots9}_{n} - 1| = 10^{-n},$$

and observe that for any positive real number ϵ we can choose a value for n that is greater than $1/\epsilon$, and such a value would satisfy $|x_n - 1| = 10^{-n} < \epsilon$. This is a formal proof that shows that the different between x_n and 1 can get as small as we want if we choose a large enough n.

> Judging Mathematics by its pragmatic value is like judging a symphony by the weight of its score.
>
> —Alexander Bogomolny.

▪ **Squaring the Circle**. The problem of squaring the circle has intrigued mathematicians, as well as mathematical crackpots, since antiquity. The Greeks did not have algebra, but were strong on geometry. They could easily draw accurate circles and straight lines, which is why they were interested in geometric problems that could be solved using just these two tools.



Figure 4: Image by MS Copilot, Aug. 2025.

The Greek philosopher Anaxagoras was among those who asked whether it was possible to construct, with the help of just a compass and straightedge, a square whose area equals the area of a given circle. The problem is worthy of consideration, because a circle is curved at every point, while a square consists of only straight segments.

A decisive answer came in 1882, when Ferdinand von Lindemann proved that π is transcendental. This implies that squaring the circle with just a compass and straightedge is impossible, because a previous result had shown that using only these tools, it is impossible to construct a segment of transcendental length.

(The ancient Greeks had geometry, which enabled them to construct complex figures with just a compass and straightedge. With these tools it is possible to do quite a lot of arithmetic. It is possible to perform the four arithmetic operations and even construct certain geometric figures whose dimensions are irrational, such as the hypotenuse of a right-angle triangle. However, it is impossible to construct figures whose dimensions are transcendental.)

> I continued to do arithmetic with my father, passing proudly through fractions to decimals. I eventually arrived at the point where so many cows ate so much grass, and tanks filled with water in so many hours. I found it quite enthralling.
>
> —Agatha Christie, *An Autobiography*.

But some stubborn people never give up, and in 1925, the eminent logician Alfred Tarski gave this ancient problem a twist by dropping the compass and straightedge and adopting a pair of scissors instead. He proposed the problem of dissecting the circle, can a circle be split or broken with just scissors into a finite number of parts that can then be assembled into a mathematically perfect square with the same area?

Such a process, which is called circle dissecting and which is similar to the well-known Tangram puzzle, requires a pair of mathematical scissors, which have infinite precision, but such details have never deterred the followers of the art of mathematics.

(Tangram is an ancient dissection puzzle consisting of seven polygons that form a square. They can be arranged and rearranged to form thousands of recognizable patterns.)

Then came a disappointment. In 1964, the trio Dubins, Hirsch, and Karush discovered that, working with only scissors, a circular disk is not congruent to any strictly convex object, and the square, obviously, is convex. In simple terms, such a dissection requires pieces that are too complex to be cut by scissors, because they resemble fractals whose edges can be infinitely jagged.

But the year 1990 brought the next step, this time positive, in the saga of circle squaring. Miklós Laczkovich managed to prove that the circle can be decomposed (not dissected but decomposed) into complex-shaped, strange pieces that can be reassembled into a square of equal area just by translating them; no rotations or reflections needed.

A dissection can be done with scissors, because the pieces required for a decomposition are too complex to be cut by scissors, not matter how delicate. Laczkovich replaced scissors with a tool known as the axiom of choice, discussed below.

His proof allows for pieces as strange as dust particles. The proof promises that such a process is possible, but it is an existence proof; it doesn't tell what the actual pieces are and how to move and assemble them. Even worse, his pieces are unmeasurable, meaning it's impossible to measure and determine their area. We also do not know how many pieces may be needed, and Laczkovich only provides an upper bound of 10^{50}, a vast number!

Nevertheless, this revolutionary proof of Laczkovich's served to renew interest in the problem and in 2006 came a publication, by Łukasz Grabowski, Andras Máthé, and Oleg Pikhurko, that is almost constructive. Most of the pieces they use are well defined, but their assembly into a square left a small hole in the square, a hole described as "a set of measure zero."

Clearly, at least one more step was needed to fully decompose the circle, and this came in 2015 in the form of work by Andrew Marx and Spencer Unger. Now there is a complete description of a method that decomposes the circle completely, but requires an even unimaginably larger number of pieces, in the order of 10^{200}.

This latest achievement may not be the last word in the long saga of squaring the circle. As long as mathematicians exist, surprises in this area, as in many others in mathematics, may await us.

[Nadis 23] is a short, nontechnical reference to the history described here.

Appendix: The Axiom of Choice

Mathematicians sometimes hear the following question from friends "What is the axiom of choice? Does it have anything to do with free choice?" Here is a nontechnical description of this strange concept and its history.

Set theory, the creation of Georg Cantor in the 1880s, was one of the biggest surprises in the history of mathematics and has remained controversial since its inception. The theory talks about the relative sizes of infinite sets and deduces in a simple way that there are different kinds, or magnitudes, of infinity.

Well, are there different infinities? The simplest answer is We don't know. The human mind cannot grasp infinity, and by definition, infinity cannot be reached and examined. A better answer is that like any theory, set theory is based on assumptions that underlie it and are responsible for its results and conclusions. Cantor's initial, naive assumption was that sets, even infinite sets, can be compared by matching their elements, and based on this assumption he arrived at his revolutionary results.

It didn't take long for cracks to develop in set theory. Already in 1901, Bertrand Russell discovered and published the paradox that is named after him and which showed that something was wrong with set theory, at least with the way Cantor had originally developed it. Admittedly, Russell's work dealt with a strange object, the set R whose elements are all the sets which are not members of themselves, but this is acceptable because mathematics is full of unusual, small and large, beautiful and ugly, objects, ideas, and proofs. Russell used basic logic to prove that if R is an element of itself, then R is not an element of itself, and conversely, if R is not an element of itself, then R is an element of itself. Expressed in symbols, this paradox says:

$$\text{Let } R = \{x | x \notin x\}, \text{then } R \in R \Leftrightarrow R \notin R.$$

Those mathematicians who objected to set theory from the beginning were not surprised by this contradiction, but the many friends and enthusiasts of the theory got concerned, and some have immediately decided to try and put it on a firm mathematical basis. A mathematical theory is normally based on axioms; simple self-evident statements on which later conclusions, results, and proofs are based. The term `axiom` comes from the Greek $\alpha\xi\iota\omega\mu\alpha$ (axiōma) meaning "that which is thought worthy or fit" or "that which commends itself as evident."

Several mathematicians helped in choosing a set of axioms on which the new Axiomatic Set Theory or Zermelo–Fraenkel set theory (ZF), would be based, chief among them were Ernst Zermelo and Abraham Fraenkel. Deciding on a set of axioms is not trivial. In addition to being few and simple, axioms have to be consistent and independent, and give rise to an entire theory that would grow from them.

> Zermelo's surname appears to have been unique to his family (after his wife's death, German telephone directories showed no one else with that name), and he gave various explanations of its origin, including that it was taken from the middle of the word Walzermelodie (waltz melody).
>
> —From the lore of mathematics.

The original set chosen by Zermelo and Fraenkel consisted of six axioms as follows: Extensibility, Union, Power set, Infinity, Replacement, and Regularity.

The axiom of choice was included in Zermelo's axiomatic set theory around 1904–1908 and is now part of the modern ZF theory. In simple words, it claims Given a set C whose elements B are nonempty disjoint sets, it is always possible to (simultaneously) choose exactly one element of each B and use those elements to construct a new set.

(The word "simultaneously" is sometimes omitted.) As long as we consider finite sets, there is no problem. In principle, it is possible go over every set B and choose an element of it, either at random or according to some rule. With infinite sets, however, it is impossible to go over all the sets B, so a rule or a function is needed to specify how to choose an element from each B.

Here are examples of infinite sets C.

• If the elements of sets B are integers, we can always choose the smallest one.

• If the elements of sets B are real intervals of the form $[a, b]$, it is possible to choose the middle element $(a + b)/2$ from each set B.

• However, if the elements of sets B are real numbers, it may be impossible to specify a rule of selection. Recall from the Introduction that a real number does not have an immediate successor or a predecessor. Given a real number a, there is no number that immediately follows or precedes a. Also, an interval of real numbers may not have a smallest, largest, or even a middle number. An example is the open interval (a, b).

The next two examples are by Bertrand Russell and are about sets with non-mathematical elements.

• If each set B consists of a pair of shoes, then it is possible to choose, say, the left shoe from each B.

• However, if each set consists of a pair of identical socks, there is no way to identify one and choose it.

In all these examples the axiom of choice insists that the answer is yes, it is possible to choose, but it doesn't tell how.

Being more philosophical than mathematical, the axiom has been a source of controversy and arguments since its inception. It has also produced unexpected and unusual consequences of which the Banach-Tarski paradox is the most spectacular. This claim has to do with the decompositions of the sphere, and is therefore related to the question of circle squaring discussed earlier. The claim is that any solid sphere can be split into finitely many pieces which can then be reassembled to form two solid spheres (solid, but unmeasurable) of the same size. Furthermore, any solid sphere can be split into finitely many pieces in such a way that they can later be reassembled to form a solid (but unmeasurable) sphere of arbitrary size.

References [veritasium 25] and [vitali 25] illustrate other important applications of the axiom of choice. The former shows how the axiom was used to prove Cantor's "well-ordering of the real numbers" theorem, while the latter was used to construct Vitali's set, a set of real numbers that can have no measure; it is unmeasurable.

The following sentence summarizes both the nature of the axiom of choice and its consequences: "The axiom of choice is so obviously true while its consequences are so obviously false, that we must ask ourselves 'what the hell is going on here'."

However, each generation brings a new crop of young and original mathematicians, and in 1938 the work of Kurt Gödel showed that there is a version of mathematics where the axiom of choice is as true as the other six ZF axioms. Later, in 1963, Paul Cohen proved that there is also a version of mathematics where the six original ZF axioms are true but the axiom of choice does not.

This is confusing, so here is an analogy. Geometry is also based on axioms. The first four axioms chosen by Euclid are concerned with points, straight lines, circles, and

right angles, but there is the fifth axiom about parallel lines never meeting. Today, we realize that there are two geometries, the one where this axiom is true and another where parallel line do meet. Two versions of geometry, or two different worlds with different mathematics.

> The universe can be very a strange place without choice. One consequence of the Axiom of Choice is that when you partition a set into disjoint nonempty parts, then the number of parts does not exceed the number of elements of the set being partitioned. This can fail without the Axiom of Choice.
>
> —The mysterious Dr. Strangechoice.

- **Tupper's Self-Referential Formula**

Tupper's self-referential formula, developed by Jeff Tupper in 2001, is a formula that visually represents itself when graphed at a specific location in the (x, y) plane. Wikipedia defines the term self reference as follows:

> Self-reference is a concept that involves referring to oneself or one's own attributes, characteristics, or actions. It can occur in language, logic, mathematics, philosophy, and other fields.
>
> In natural or formal languages, self-reference occurs when a sentence, idea, or formula refers to itself. The reference may be expressed either directly, through some intermediate sentence or formula, or by means of some encoding.

The following software challenge has long been a favorite of aficionados of computer programming (software design or software engineering). It asks you to write a program that prints itself. Being nerds, many programmers love this kind of challenge and have spent much time and effort in creating this type of, basically useless, software. (Useless, because it requires no input. It always outputs the same thing. See reference [Quine 23].)

This kind of self-replicating of self-producing software has been named Quine by Douglas Hofstadter in honor of the philosopher Willard Quine.

At first, Tupper's formula [Tupper 01] seems to be self-referential, but once we understand how it works and what it really does, it becomes clear that this formula is something different; it is more a fun game that connects binary numbers and monochromatics (black-and-white) graphics. It is general and can print more than just itself, and it is based on the frightening-looking inequality

$$\frac{1}{2} < \left\lfloor \mathrm{mod}\big(\lfloor \frac{y}{17} \rfloor 2^{-17\lfloor x\rfloor - \mathrm{mod}(\lfloor y\rfloor, 17)}, 2\big)\right\rfloor . \tag{.1}$$

To draw a figure using this inequality, prepare a $[107, 18]$ bit rectangle and fill it with zeros. Scan it by varying x in $[0, 106]$ and y in $[k, k + 17]$, where k is the 543-digit ($1802 = 106 \times 17$ bits,) integer listed in Figure 5. For each pair of coordinates (x, y) that satisfies the inequality above, set the corresponding bit in the grid to 1. The low-resolution figure obtained in this way will represent the inequality.

Even more. Take any image, scan it in monochrome and in any low resolution, not just 107×18, to obtain a grid of bits, a 1-bit for a black pixel and a 0-bit for a white pixel. Scan this grid column by column, collecting the bits. The resulting bitstring can then be used as k in Figure (.1).

```
ArrayPlot[Table[
 Boole[1/2<Floor[Mod[Floor[y/17] 2^(-17 Floor[x]-Mod[Floor[y],17]),2]]],
  {y,k,k+16},{x,105,0,-1}]/.
k->
960939379918958884971672962127852754715004339660129306651505519271702 8
023952664246896428421743507181212671537827706233559932372808741443078 9
132596394133772348785773574982392662971551717371699516523289053822161 2
403238855866184013235585136048828693337902491454229288667081096184496 0
917051834540678277315517054053816273809676025656250169814820834187831 6
384911559022561000365235137034387446184837873723819822484986346503315 9
410054974700593138339226497249461751545728366702369745461014655997933 7
985374831437868418065934222278983887229800007484047 19]
```

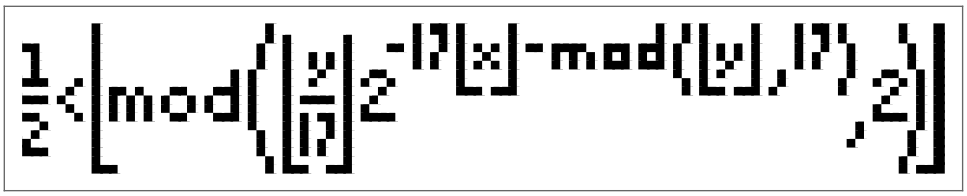


Figure 5: Tupper's Self-Referential Formula and Its Output.

With this process in mind, the Tupper formula is no longer magical or even complicated. It is only part of a straightforward process of converting a monochrome image to a number and converting any number to such an image. We can think of it as either (1) encoding an image in a number, in cases where it is easy to share a number than an image, or (2) a way to hide a number in an admittedly weird looking, but innocuous image.

To convert an integer to an image, it should be converted to a bitstring and arranged in a rectangular array. Figure 6 shows, on lines 2 through 4, such an array, of dimensions 8×10. The array is displayed in matrix form on line 5. An empty list `l` is defined on line 7 and is populated, on lines 8 and 9, with the pairs (x, y) of coordinates of array locations that have a 1-bit. The `Show` command on lines 12–13 then uses the coordinates to display black rectangles that recreate the original image.

The Tupper formula of Figure (.1)is more complex than the simple test on line 8, because it creates from each coordinate pair (x, y) a 16×16 square of pixels. This is done in order to end up with large pixels and an easy-to-read final image.

- **Ramsey Theory: complete chaos is impossible**

Everyone knows that the pope is the head of the Roman Catholic Church. Fewer people know that the bishop of Canterbury is the primate of the Anglican church (or Church of England). Even fewer have heard the name of Michael Ramsey, the 100th bishop of Canterbury (1961–1974). Practically no one outside of a small circle of mathematicians have heard of his brother, Frank who died aged 26, and yet it is Frank Ramsey who developed one of the most original mathematical fields, known today as Ramsey theory. It is said that Frank Ramsey was interested in English literature, the classics, politics, philosophy, and psychoanalysis, while also making important contributions to mathematics. This short discussion of Ramsey theory follows reference [ram-

```
Remove["Global`*"];
k={{1,1,1,0,1,1,1,0,0,0},{1,0,0,0,0,1,0,0,0,0},{1,1,1,0,0,1,0,1,0,1},
 {1,0,0,0,0,1,0,1,1,1},{1,0,0,0,1,1,1,0,1,0},{0,0,1,0,0 ,0,0,0,0,0},
 {0,1,0,0,1,1,1,0,1,0},{1,0,0,1,1,0,0,1,0,1}};
MatrixForm[k] (* k is 8x10*)
r=First[Dimensions[k]]; c=Last[Dimensions[k]];
l={};
Do[{x=Floor[(i-1)/c+1],y=Mod[i-1,c]+1,If[k[[x,y]]==1,AppendTo[l,{x,y}]]},
 {i,1,r *c}];
len=First[Dimensions[l]];
bbox:={EdgeForm[Thick],Rectangle[{0,0},{.8,.8}]};
Show[Graphics[{Table[Translate[ bbox,{l[[t]][[2]],-l[[t]][[1]]}],
 {t,1,len}]},Background->LightBlue]]
```

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

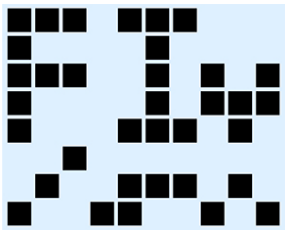


Figure 6: Mathematica Code for Tupper's Formula.

sey.bazett 23].

Ramsey theory is part of combinatorics, a vast field which is concerned with the enumeration (counting) of mathematical structures, arrangements, or configurations. However, Ramsey theory is different from the rest of combinatorics because it studies the amount of order in a structure as a function of its size. A topic in Ramsey theory may ask, given this structure, how can we prove that it, or a bigger version of it, would satisfy a particular property or a given relation. Sometimes, the aim is more specific, and we ask how big should the structure be in order to have the desired property or relation? The following are two key theorems of Ramsey theory:

- Van der Waerden's theorem. For any given c and n, there is a number V, such that if V consecutive numbers are colored with c different colors, then the sequence of numbers must contain an arithmetic progression of length n whose elements are all the same color. Here is a typical example. Given a sequence of eight numbers, each colored with one of two colors R and B, we check for the existence of arithmetic progressions.

  It is easy to verify that for the particular coloring $R_1\ B_2\ B_3\ R_4\ R_5\ B_6\ B_7\ R_8$ there are no arithmetic progressions. As we start at a particular color and hop from left to right in equal steps, we cannot reach the same color more than once. However, adding a ninth cell of any color, R or B, makes it possible to move, for example, is steps of 4, from 1 to 5 to 9 (if 9 has color A) or in steps of 3, from 3 to 6 to 9 (if 9 has color B). By extending the structure from 8 to 9 units, we have opened up possibilities of skipping in equal steps while staying with the same color.

- The more general Hales-Jewett is an existence theorem. It guarantees that something must exist, but cannot tell us how to construct it or how big it must be. Specifically,

the theorem says that an object with a high enough number of dimensions would feature some combinatorial requirement. In other words, just by increasing the number of dimensions of the object, we can always eliminate any element of randomness that it originally exhibited. Some may even draw from this the extreme conclusion that there is no such thing as complete randomness. Stated informally, the theorem says that for any given n and c, there is a number H such that if the cells of an H-dimensional $n \times n \times n \times \cdots n \times n$ cube are colored with c colors, there must be one row, column, or diagonal all of whose cells are the same color. The coloring cannot be completely random.

The following example of a graph illustrates the Hales–Jewett theorem. A mathematical graph, such as the one of Figure 7a, is a set of nodes, also called vertices, connected by edges. The 6-vertex graph in the figure describes a group of six people who happen to be in the same place (Ramsey used the example of a party). Some of them may know certain others, but perhaps not everyone. The blue edges in the figure connect people who know each other, while the red edges connect pair of people who don't know each other. It is clear that the blue edges, which form a subgraph, also define several triangles, each of which is a small group of three people where each one knowns everyone else. The same is true for the red edges. The Hales–Jewett theorem states that regardless of how many people each of the six people knows, there will always be at least one triangle, blue or red, in a group of six people, and more triangles can be expected with larger parties.

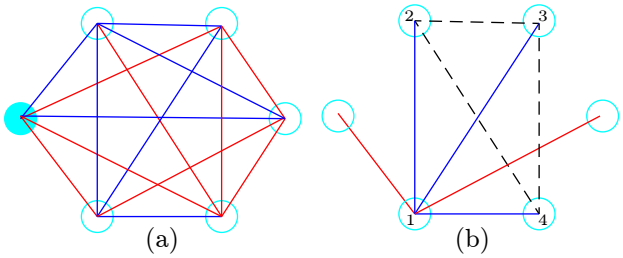


Figure 7: Two Fully-Connected Graphs.

The proof is simple. The full circle in Figure 7a is a person connected to five other people, some of which he knows. Obviously, given that there are only two colors, at least three of the five edges emanating from this person must have the same color. Let's assume that the three same-color edges of vertex 1 in Figure 7b are blue. They lead from vertex 1 to vertices 2, 3, and 4, so we connect those vertices with dashed lines and observe the following. If all three dashed lines are red, then there is a red triangle in the graph. Otherwise, at least one of the three must be blue, and whichever it is, it forms a blue triangle between vertex 1 and one of the other numbered three vertices. Thus, there will always be at least one triangle in a graph of six or more vertices.

Figure 8 is another example of this theorem. It is easy to verify that none of the rows and diagonals of the two-dimensional tic-tac-toe boards is uniform, but once these are extended to a three-dimensional structure, there are uniform substructures, two of which are shown in red dashed lines.

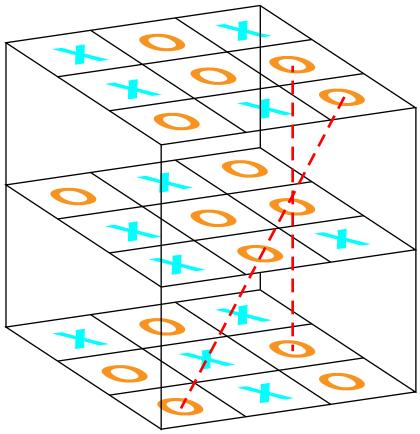


Figure 8: A 3D Tic-Tac-Toe.

> There are only two kinds of math books: Those you cannot read beyond the first sentence, and those you cannot read beyond the first page.
>
> —Chen-Ning Yang.

■ **The birthday Problem (or Paradox)**. This interesting problem, with its unexpected and counterintuitive result, is an example of the use of probabilities to gain an insight into problems. We present a simple, approximate solution and illustrate it with examples. For a more detailed treatment, see reference [Wikibirthday 23] and Section 7.9.

Imagine n people in a room, perhaps for a birthday party. Each has a birthday (just month and day, not year). The birthday problem is to compute the probability that two or more of them share the same birthday.

This probability is unexpected and is often referred to as a paradox, because it is higher than what most people feel intuitively. For example, the case $n = 23$ (just 23 people with random birthdates) would cause many people to guess a low probability of sharing a birthday, but the correct (albeit approximate) result is about 50%. With $n = 50$, this probability rises to about 97%.

Here is one way to compute this probability. We ignore leap years, we assume that all days are equally likely as birthdays, and that there are no twins in the room.

We start with person #1. He/she was certainly born, so the probability of him having a birthday is 1. We denote his birthday by D and decide to write as 365/365. The probability of person #2 NOT sharing birthday D is 364/365, because he must have been born on one of the other 364 days. We denote his birthday by F. Similarly, for person #3 not to share his birthday with either #1 or #2, he must have been born not on day D and not on day F, but on one of the remaining 363 days. His probability of not sharing a birthday is therefore 363/365.

This argument is similarly extended to all the remaining persons. The probability of person #n not to share his birthday with the $n-1$ others is, similarly, $(365-n+1)/365$.



Figure 9: Image by chatGPT 4o, Aug. 2025.

Thus, the probability that no two people share a birthday is the product

$$\begin{aligned}
&\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{(365-n+1)}{365} \\
&= 365 \times 364 \times 363 \times \cdots \times (365-n+1) \left[\frac{1}{365}\right]^n \\
&= \frac{365!}{(365-n)!365^n} \\
&= \frac{n! \cdot \binom{365}{n}}{365^n},
\end{aligned}$$

where symbol `!` is the factorial operator and $\binom{365}{n}$ is the well-known binomial coefficient. The Mathematica code

```
n = 23.;
n! Binomial[365, n]/365^n
```

produces 0.492703. Subtracting this from 1 yields 0.507297. Surprisingly, the probability of 23 people, chosen at random, to not share a birthday, is higher than 50%. Similarly, setting $n = 50$ results in 0.0296264, which produces the probability $1 - 0.0296264 = 0.9703736$.

Such a computation can be viewed as the product of evaluating each person against the other $n-1$ persons. Clearly, it is not enough to check whether persons 2 through n share a birthday with person 1. We also have to check whether persons 3 through n share a birthday with person 2, whether persons 4 through n share a birthday with person 3, and so on.

An interesting, practical application of this result in probability is implementing safe passwords. When a computer user opens a new account at, say, a location X, and sends in his new password for the account, the computer at X passes the password

through a simple hashing process where the password's characters are scrambled and put through several calculations to produce a fixed length hash code C, normally 128 or 256 bits long, which is then stored in computer X.

The password itself is then deleted. It is not saved anywhere in location X, to prevent it from being stolen. The hashing process is irreversible, which means that even if the table of hash codes is stolen, the codes cannot be converted back into passwords.

Each time the user logs into his account at X and types the password, the computer at X hashes it, computes a hash code D, and then searches the table of hash codes for a match to D.

This process results in high digital security at location X, but it requires that no two passwords would hash to the same hash code. In professional terms, it is mandatory to have no collisions. This is where the birthday paradox is employed to maximum effect. The equivalent of persons is the passwords. The equivalent of the birthdays are the new users inputting passwords. That of 365 is the size of the hash codes space, and the equivalent of n is the current number of users.

With 128-bit hash codes, there can be $2^{128} = 3.40282 \cdot 10^{38}$ codes, resulting in a space much bigger than any conceivable number of passwords. With 256-bit hash codes, the size of the space is $1.15792 \cdot 10^{77}$, unimaginably vast.

This book has been a journey,...
—Paul R. Walker.

1
Graphics: Visible Math Objects

A common definition of art in a dictionary is "the creation of work or its principles; the act of making or producing things that exhibit form, beauty, and unique perspective." In English, the term art also has various other connotations like the skill of negotiation, humanities, and cunning behavior. What catches our attention is the link established in this definition between art and beauty. An artwork must possess beauty regardless of its creation process or inspiration. Therefore, the mathematical principles utilized in creating art are connected with aesthetics.

This chapter is concerned with those aspects of mathematics that can be employed to create beautiful or attractive patterns, or simply patterns that are pleasing to the eye. One of the marvels of mathematics is the pleasure we feel when an unexceptional, dry-looking mathematical expression unexpectedly produces a lovely geometric shape when plotted graphically.

Those with a mathematical background know that mathematical functions and expressions can be plotted graphically as curves or surfaces, many of which would be considered interesting, nice, lovely, attractive, pretty, fair, or even beautiful. Some examples of such plots are shown in this chapter, but even more interesting graphical objects can be constructed from other approaches such as: (1) combining straight lines or arcs in various ways, (2) plotting the curve followed by an object when it slides without friction along the circumference of another object, (3) using random numbers to combine familiar objects in unpredictable, surprising ways, and (4) extending two-dimensional figures to three dimensions, such as constructing a helix from a spiral. These and other approaches are described and illustrated in this chapter.

1.1 Curves and Surfaces

We start with simple functions. In English, the word *function* has several meanings, one of which is a mathematical function. A mathematical function is a rule of computation. Given a number x, the function $y = f(x)$ specifies how to compute another number y. Consider, for example, the function $f(x) = x^2$. For $x = 1$, the function specifies (returns) $y = 1^2 = 1$. For $x = 2.1$, the function returns $2.1^2 = 4.41$, and so on. Applying this function to all the real numbers x in the interval $[1, 2]$ returns all the real numbers y in the interval $[1, 4]$. We can therefore say that a function produces pairs (x, y) of numbers.

The discipline of analytic geometry (the study of geometry using a coordinate system, also known as coordinate geometry or Cartesian geometry) teaches us that such a pair specifies a point in two dimensions. The first component of the pair is the x coordinate of the point; it specifies its distance from the y axis. Similarly, the second component is the y coordinate of the point, i.e., its distance from the x axis (Figure 1.1a). It is now easy to see how the set of all the pairs (x, y) specified by a function can be plotted as a set of points to generate a curve. Figure 1.1b shows the curve of the function $y = x^4$ for x values in the interval $[1, 2]$ and Figure 1.1c shows a sine curve from 0 to π. Part (d) of the figure shows how such smooth curves are drawn in practice. Several points on the curve are computed and are connected with straight segments. If the segments are short enough, the curve appears smooth.

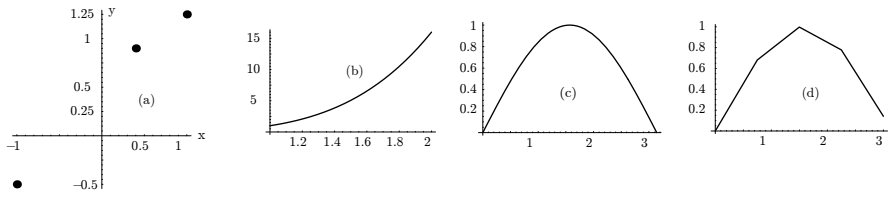


Figure 1.1: Points and Curves.

> I made a graph of all the past relationships I had. It had an "ex" axis and a "why" axis.

Curves can also be three-dimensional, as the helix of Figure 1.2a illustrates. A function may also specify how to compute a result z from two given numbers (parameters) x and y. Such a two-parameter function has the form $z = f(x, y)$ and it can be employed to plot a surface. For any pair (x, y), the function computes a third number z and the triplet (x, y, z) is plotted as a point in three dimensions. Once enough points have been plotted, the result looks like a surface. Figures 1.2b,c show the surface produced by the functions $\cos(x \cdot y)$ and $e^{-(x^2+y^2)}$.

A function of the form $y = f(x)$ is called explicit because it separates the two variables x and y. A function where the two are not separated is called implicit and has the form $F(x, y) = 0$. The explicit equation of a straight line is $y = ax + b$ and that of a sine curve is $y = \sin(t)$. However, a circle is easier to express implicitly as $x^2+y^2-R^2 = 0$. (Many other closed curves also lend themselves to implicit representations.)

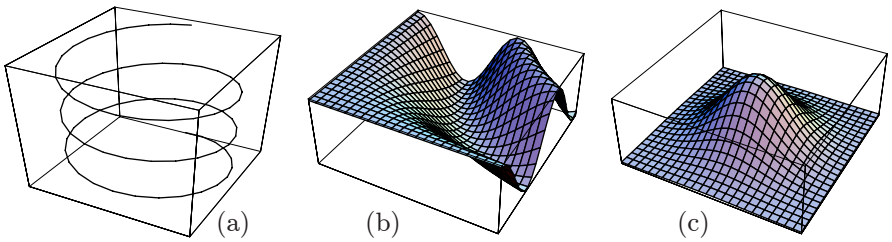


Figure 1.2: Three-Dimensional Plots.

A function can also be *parametric*. A parametric representation is an approach to curve and surface design where attractive curves and surfaces can be created systematically and made as complex as desired. The idea is to represent the coordinates of points on the curve (or surface) as functions of one (or two) parameters. Thus, the expression $(\cos(t), \sin(t))$ is a two-dimensional point whose x and y coordinates are functions of a parameter t. When t is varied in a certain interval, a point $(\cos(t), \sin(t))$ is computed for each value of t and the total of all the points becomes a curve. The curve generated by $(\cos(t), \sin(t))$ is a circle because $\sin^2(t) + \cos^2(t) = 1$ for any t. This basic circle can be extended to an ellipse, a spiral, an oval, and many other curves. It can also be extended to three dimensions to become a helix, a cylinder, a cone, and other surfaces and space curves.

The three-dimensional helix curve of Figure 1.2 was computed by the parametric function $\big(\sin(3t), \cos(3t), t/3\big)$ for t in the interval $[0, 2\pi]$.

A parametric surface is generated by a function of the form $\big(x(u,w), y(u,w), z(u,w)\big)$ when it is computed for many values of the parameters u and w and the resulting points are plotted. Thus, for example, the function $(u, w, 3)$ corresponds to a flat plane at $z = 3$. Here are examples of parametric surfaces that are attractive, exhibit certain symmetry, and are easy to compute, display, and extend in many ways.

1. A saddle (formally, a hyperbolic paraboloid) has the form $z = x^2 - y^2$, and can therefore be computed parametrically by $(u, w, u^2 - w^2)$ when both u and w are varied symmetrically about zero.

2. The explicit expression $z = \cos(x^2 + y^2)$ corresponds to a wavy surface. It is easy to represent parametrically as $\big(u, w, \cos(u^2 + w^2)\big)$.

3. A cylinder is a circle extruded vertically, which suggests the parametric expression $\big(\cos(u), \sin(u), w\big)$ where u varies in the interval $[0, 2\pi]$ and w goes from 0 to the desired height of the cylinder.

4. A disk is a surface that consists of concentric circles located at the same height. Its parametric representation is therefore $\big(w\cos(u), w\sin(u), 0\big)$.

5. A cone is the surface obtained when the same concentric circles are located higher and higher on the z axis. Instead of being always 0, the z coordinate of a circle is slightly larger than that of the preceding circle. The parametric expression is therefore $\big(w\cos(u), w\sin(u), w\big)$. A double cone can be obtained when w is varied in an interval that is symmetric about 0, such as $[-2, 2]$.

Figure 1.3 illustrates the next five surface constructions. They are all based on circles of radius w whose heights are varied in simple ways.

6. The first surface of this quintet is a flower. The idea is to vary the height

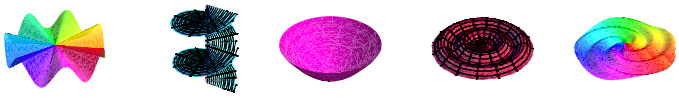


Figure 1.3: Five Parametric Surfaces.

of each circle periodically while the circle is plotted. This is achieved by the simple parametric expression $\big(w\cos(u), w\sin(u), (w\cos(7u)/2)/2\big)$. Notice how the height (the z coordinate) is varied as $\cos(7u)$.

7. When the height of a circle is proportional to u, the result is a helicoid. Each curve on this surface is a helix and is obtained by drawing a circle and raising it continuously as it is being drawn. The parametric expression is $\big(w\cos(u), w\sin(u), 0.7u\big)$.

8. A paraboloid is obtained when the height of a circle is equal (or proportional to) the square of its radius. The expression is $\big(w\cos(u), w\sin(u), w^2\big)$.

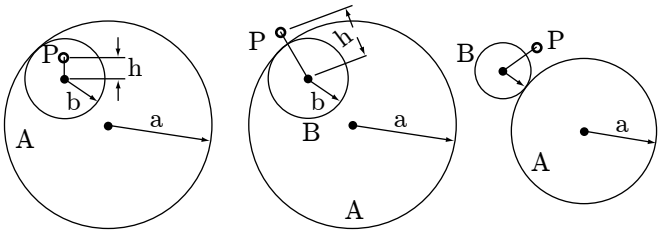
9. When we vary the height of a circle as the sine or cosine of its radius, the result is a strange surface of revolution, somewhat resembling a torus. The surface in the figure was generated by $\big(w\cos(u), w\sin(u), \cos(w)\big)$. Note how consecutive circles have larger and larger radii, followed by circles of smaller and smaller radii.

10. A strange, unnameable surface results when the height of a circle varies as its radius w, but is also varied while the circle is being drawn. The surface in the figure was computed by $\big(w\cos(u), w\sin(u), \cos(w-4u)\big)$.

◇ **Exercise 1.1:** Try to guess the shape of the surface produced by the parametric expression $\big(w, p\cos(u), p\sin(u)\big)$ where $p = 1+\cos(w)$.

1.1.1 Hypotrochoids, a Family of Curves

Given a circle A of radius a, we place a smaller circle B of radius b inside it and let B roll without sliding along the circumference of A. A point P is attached to circle B at a distance h from its center. The roulette curve traced by P is called a hypotrochoid. The shape of the curve depends on a, b, and h, which is why the hypotrochoids constitute a family of curves, many of which are well known and have their own names. Spirograph is a toy that employs plastic sprocket wheels to trace many hypotrochoids. If circle B rolls along the circumference of A, but *outside* A, the curve traced by P is referred to as epitrochoid.



It can be shown that the parametric expression for a general hypotrochoid is

$$H_{a,b,h}(t) = \left[(a-b)\cos(t) + h\cos\left[\frac{(a-b)t}{b}\right], (a-b)\sin(t) - h\sin\left[\frac{(a-b)t}{b}\right]\right],$$

and the general epitrochoid is similarly expressed by

$$E_{a,b,h}(t) = \left[(a+b)\cos(t) - h\cos\left[\frac{(a+b)t}{b}\right], (a+b)\sin(t) - h\sin\left[\frac{(a+b)t}{b}\right]\right].$$

There are many special cases of hypotrochoids, among which should be mentioned the hypocycloid (the case where $h = b$), the ellipse (when $a = 2b$ and $h \neq 0$), and the n-petal rose (obtained for $a = 2nh/(n+1)$ and $b = (n-1)h/(n+1)$. Figure 1.4 shows six rows of seven hypotrochoids each. Each row corresponds to a different pair (a, b) and the curves of a row correspond to h values of 0,5, 1, 1.5, 2, 2.5, 3, and 3.5.

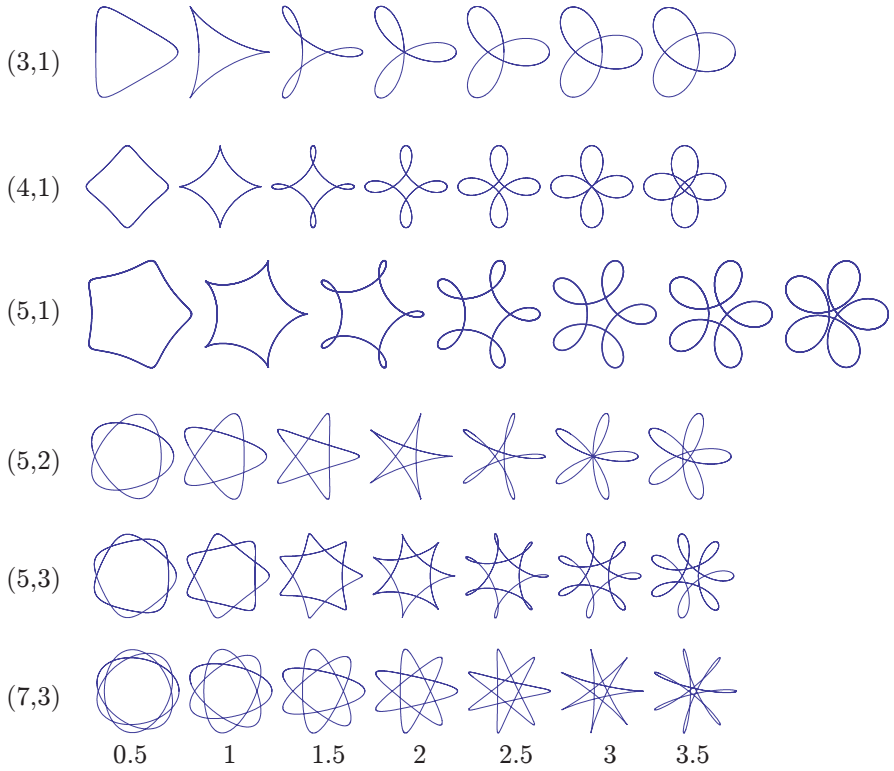


Figure 1.4: Various Hypotrochoids.

It is easy to see how the shapes vary from a simple triangle to roses with seven petals.

A cycloid is the curve traced by a point on a circle as it rolls along a straight line without slipping (Figure 1.5a). A cycloid is a special form of trochoid and is an example of a roulette, a curve generated by a curve rolling on another curve. Part (b) of the figure shows how the cycloid is also the curve of shortest time. Given points A (higher) and B (lower) in a field of uniform gravity, we want to connect them with a curve such

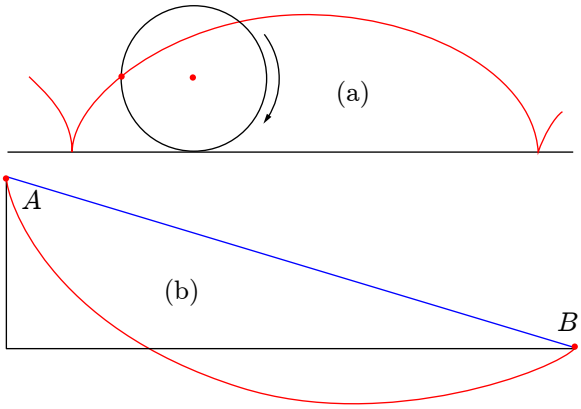


Figure 1.5: Cycloid and Brachistochrone.

that an object sliding down it without friction would take the shortest time. Such a curve is called a brachistochrone (from Greek: "shortest time").

Given the cycloid that is generated by a circle of radius r rolling over the x-axis on the positive y side, its parametric equation is given by $\big(r(t-\sin t), r(1-\cos t)\big)$, where t is a real parameter corresponding to the angle through which the rolling circle has rotated.

> A mathematician is a device for turning coffee into theorems.
> American coffee is good for lemmas.
> —Alfréd Rényi and Paul Erdős.

◇ **Exercise 1.2:** Use the many examples of reference [sine.amazing 24] to experiment with simple combinations of sine and cosine curves that create pleasing patterns, especially with straight lines.

1.1.2 Patterns From Straight Segments

A straight line is not generally considered an object of beauty, but it is possible, even easy, to obtain beautiful patterns from groups of straight segments where the individual segments partially overlap or vary their orientations in simple ways.

The first idea is to select a long, preferably winding, curve $P(t) = \big(x(t), y(t)\big)$, select a constant a, and generate a single polyline by connecting points of the form $P(an)$ and $P(a(n+1))$ for $n = 0, 1, 2, \ldots$. If the resulting pattern is not interesting, a new, perhaps very different pattern may be generated by selecting another constant a.

The first example is a circle. One parametric expression of a circle is $\big(\cos(t), \sin(t)\big)$. We generate a polyline of 24 straight segments from point $\big(\cos(na), \sin(na)\big)$ to point $\big(\cos((n+1)a), \sin((n+1)a)\big)$, while varying n from 0 to 25. Figure 1.6 shows seven of the many patterns created when parameter a is varied between 0 and π.

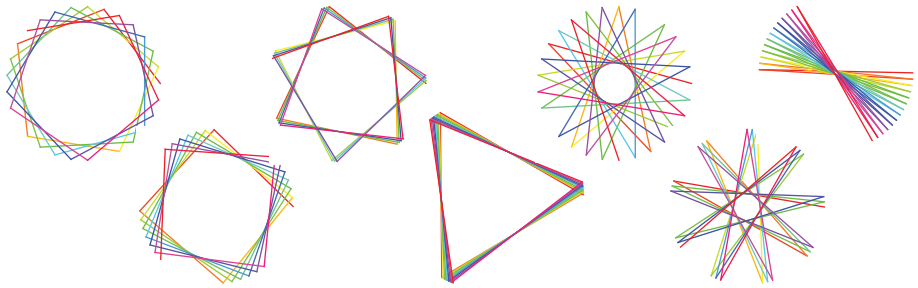
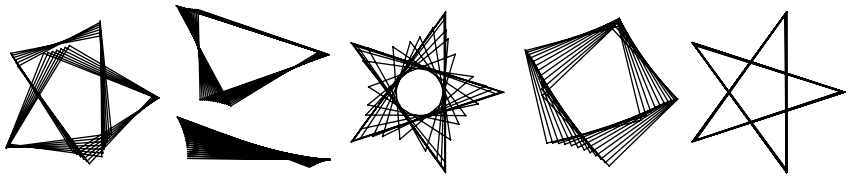


Figure 1.6: Seven Patterns Made of Straight Segments.

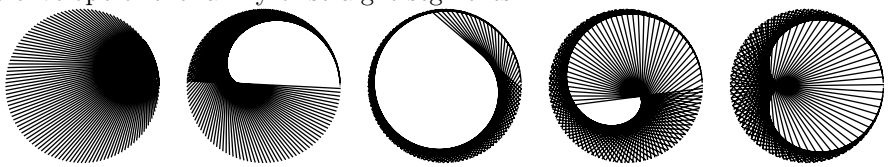
The next example starts with the 5-sided hypotrochoid $(5, 1, 1)$ (Figure 1.4), whose simplified parametric expression is $\big(4\cos(t) + \cos(4t), 4\sin(t) - \sin(4t)\big)$. The following figures were obtained by varying parameter a from 0 to 5π.



The next idea is to start with a parametric curve $P(t) = \big(x(t), y(t)\big)$, select a parameter a, connect points $P(t)$ and $P(at)$ with a straight segment, and then increment t by a small step and repeat. The five figures were generated in a circle by the Mathematica code

```
Manipulate[Show[Graphics[Table[Line[{{Cos[t],Sin[t]},{Cos[a.t],Sin[a.t]}}],{t,0,2Pi,0.05}]]],{a,0.1,2}]
```

by varying a from 0.1 to 2. The last figure (for $a = 2$) is called a cardioid and is obtained as the envelope of the family of straight segments.

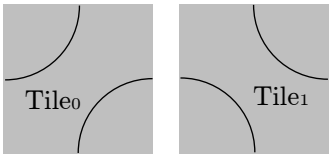


The ideas described here can be applied to any parametric curves and almost always result in a surprising variety of patterns and envelopes.

1.1.3 Patterns Formed From Circular Arcs

Here are simple examples of *tiling* the plane. Much as a floor or a wall can be tiled with vinyl, ceramic, or other types of tiles, the two-dimensional xy plane can be tiled with all kinds of patterns. The differences between real tiles and mathematical tiles are (1) the price, (2) the amount of work required to tile, and (3) the number of dimensions. In mathematics, tiles (and other objects) can have any number of dimensions, although we cannot visualize more than three dimensions.

We start with a pattern tile_0 of two circular arcs on a square and follow with a similar pattern tile_1.



The two patterns are selected at random and are arranged in a 10×10 grid to form the left part of Figure 1.7. The center part is similarly formed by randomly selecting and packing patterns of quarter disks.



Figure 1.7: Patterns of Circular Arcs and Disks.

◇ **Exercise 1.3:** The rightmost part of Figure 1.7 is a grid where a simple square pattern is copied 100 times, rotated by a random multiple of 90° each time, and packed in 10 rows and 10 columns. Try to guess the basic pattern and construct similar grids.

1.1.4 Parabola

The parabola (from the Greek $\pi\alpha\rho\alpha\beta o\lambda\eta$) is a well-known curve that often appears in nature. It is usually introduced as a conic section, but here we employ a different approach. We start with a line (called a directrix) and a point (the focus) not on the line. The parabola defined by these two elements is the locus of all the points that are equidistant from the line and the focus. Figure 1.8a shows four such points that satisfy $FP_i = P_iA_i$. A locus a curve or other figure formed by all the points whose coordinates are related mathematically. Alternatively, a locus can also be a point, line, or surface whose movement in space obeys a rigorous mathematical relation.

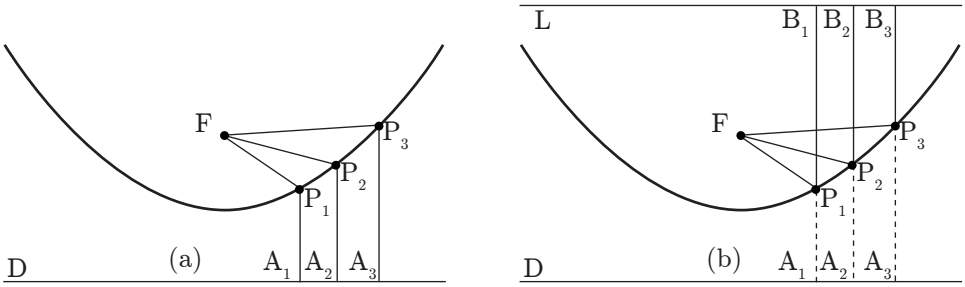


Figure 1.8: Parabolas.

Most modern telescopes are reflectors, a type invented by Newton in 1668, where the main part is a concave mirror that reflects the light and concentrates it at its focus (Figure 1.9). We intuitively feel that such a mirror should be spherical, but it is easy to show that its ideal shape is parabolic. Here is why.

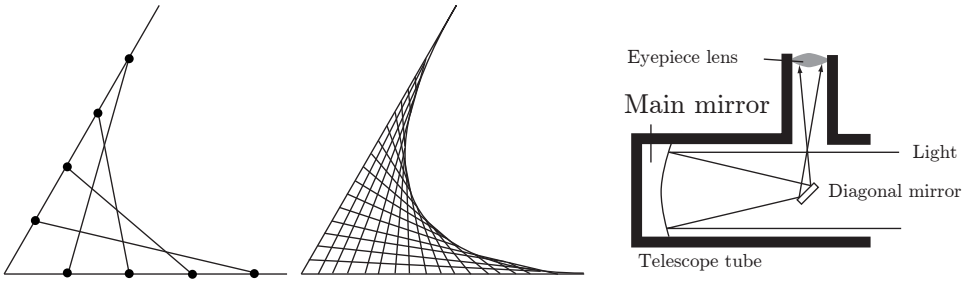


Figure 1.9: Parabolas.

Figure 1.8b shows a group of parallel light rays arriving from a star to the vicinity of such a mirror. The rays left the star simultaneously and therefore arrive at points B_i on line L at the same time. They continue to the parabolic mirror, hit it at different times (the ray from B_3 is first), and are reflected to the focus. The distance C between line L and the directrix is the same for all rays. Thus, $B_iP_i + P_iA_i = C$, but the definition of a parabola tells us that $FP_i = P_iA_i$, so we end up with $B_iP_i + P_iF = C$. All the paths of rays from line L to the focus have the same length C.

In addition to telescopes, the mirrors used in solar collectors, microwave ovens, and satellite dish antennas are parabolic (more accurate, they are paraboloids). The trajectories of comets are parabolas, as are also the path of a bouncing ball and the arc described by a stream of water sprayed at an angle from a sprinkler.

Figure 1.9 also illustrates a simple way to draw a parabola. Start with two lines that meet at a point. Place equidistant points on each line, and connect corresponding points with straight segments. The parabola is obtained as the envelope of all the segments.

Figure 1.10 illustrates how the envelope of a parabola can be obtained by folding a sheet of paper (origami). Simply fold a corner of a sheet of paper several times such that one edge of the paper always passes through the focus.

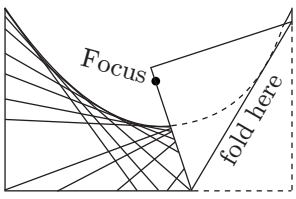
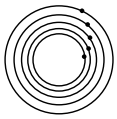


Figure 1.10: Parabola by Paper Folding.

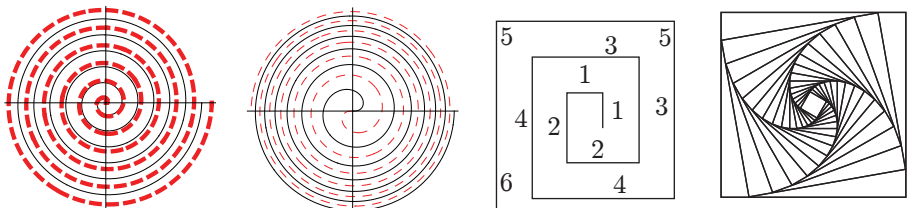
1.1.5 Spirals and Helices

Given a circle, we imagine a point moving along its circumference. While the point is moving, we continuously increase the radius of the circle. The curve generated by the point is a spiral. The precise shape of the spiral depends on the relation between the speed of the point and the speed with which the radius is increased. The special shape of a spiral implies that the most natural way to represent it mathematically is in polar coordinates. A spiral is any curve of the form $r = af(b\theta)$ where f is a monotonic increasing function and a and b are parameters.



The simplest type of this curve is the spiral of Archimedes. In this curve, the radius is increased by the same distance the point has moved. The polar equation of the Archimedean spiral is therefore $r = a\theta$, where a is a constant. After one lap (i.e., when θ is increased by 360°), the new radius is $a(\theta + 2\pi) = a\theta + 2a\pi = r + 2a\pi$. Thus, the distance between consecutive laps of this spiral is constant. The parametric equation of this spiral is obtained directly from its polar equation as $\big(r\cos\theta, r\sin\theta\big)$ or $a\theta\big(\cos\theta, \sin\theta\big)$.

The next important type of spiral is the equiangular spiral (also known as the logarithmic or Bernoulli's spiral). Its polar equation is $r = ae^{b\theta}$, where a and b are constants, leading to the parametric representation $e^{\theta}\big(\cos\theta, \sin\theta\big)$. This type of spiral has an interesting property. If we increase θ by a constant c, the radius r is multiplied by e^c. This is because $e^{(c+\theta)} = e^c e^{\theta} = e^c r$. Thus, an arithmetic sequence of θ values, results in a geometric sequence of r values. The growth patterns of the Archimedean and logarithmic spirals are compared in the leftmost part of the following figure.



Many other types of spirals are possible and have polar equations of the form $r = af(b\theta)$ for different functions f. If f is monotonic increasing, the spiral grows from the origin when θ is increased. Examples of this type are $r = \theta^2$, $r = \log\theta$, and $r = \pm\sqrt{\theta}$ (the latter rule results in the second part of the figure showing two spirals spiraling into and out of the origin). If f decreases monotonically, the curve spirals toward the origin while θ is increased. Examples of this type are $r = 1/\theta$, $r = 1/\sqrt{\theta}$, and $r = e^{-a\theta}$.

The third part of the figure shows a square spiral where two consecutive edges have the same length, and then the length is increased by one unit for the next pair. The rightmost part is a set of squares, each rotated by 10° relative to its predecessors and is shrunk by a factor of 0.86. The corners of the squares lie along four logarithmic spirals. This is an example of a pursuit curve.

A pursuit curve is the trace an object leaves when chasing another object. Imagine

four bugs located at the corners of a square. At a certain moment, each bug starts crawling towards its neighbor on the right with uniform speed. The curve followed by each bug is identical to the pursuit curves of the figure.

The two-dimensional spiral curve can be extended to three dimensions by "pulling" it out of the xy plane while it spins and expands. The three-dimensional parametric representation of the Archimedean spiral is of the form $\big(r\cos\theta, r\sin\theta, \theta\big)$. The z coordinate can either be θ or a monotonic increasing function of θ. Such an expanding three-dimensional spiral is known as a conical helix. A helix is a three-dimensional spiral whose radius is constant; it looks like a corkscrew or spiral staircase.

Figure 1.11 shows a basic helix and two three-dimensional spirals, Archimedean and logarithmic.

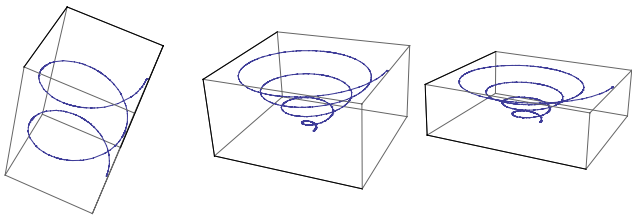
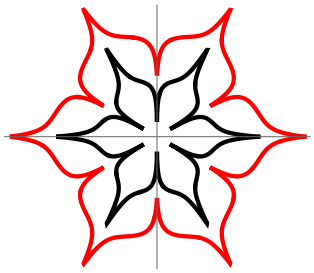


Figure 1.11: Helices.

Figure 1.12 illustrates three patterns, spirals and rosettes, based on ϕ, π, and e. Mathematica code is included. See also Figures 5.6 and 5.7.

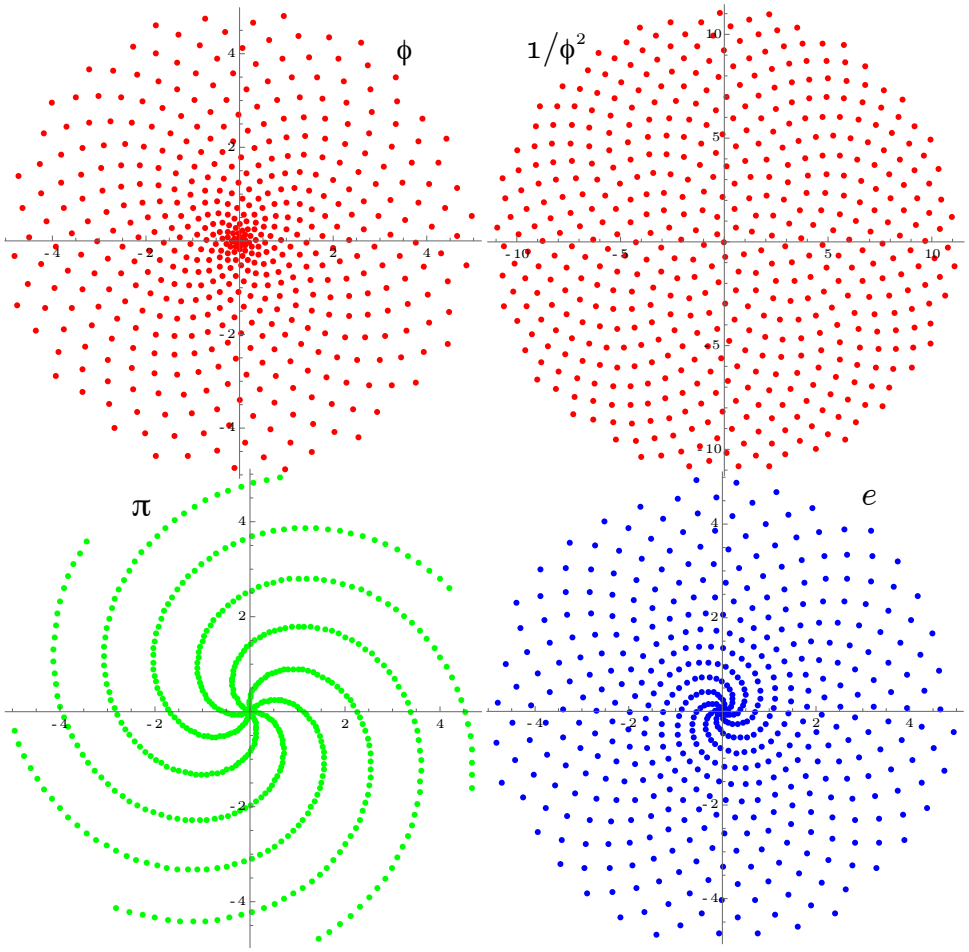
> Each generation has its few great mathematicians... and [the others'] research harms no one.
>
> —Alfred W. Adler.

1.1.6 The Lotus in Polar Coordinates



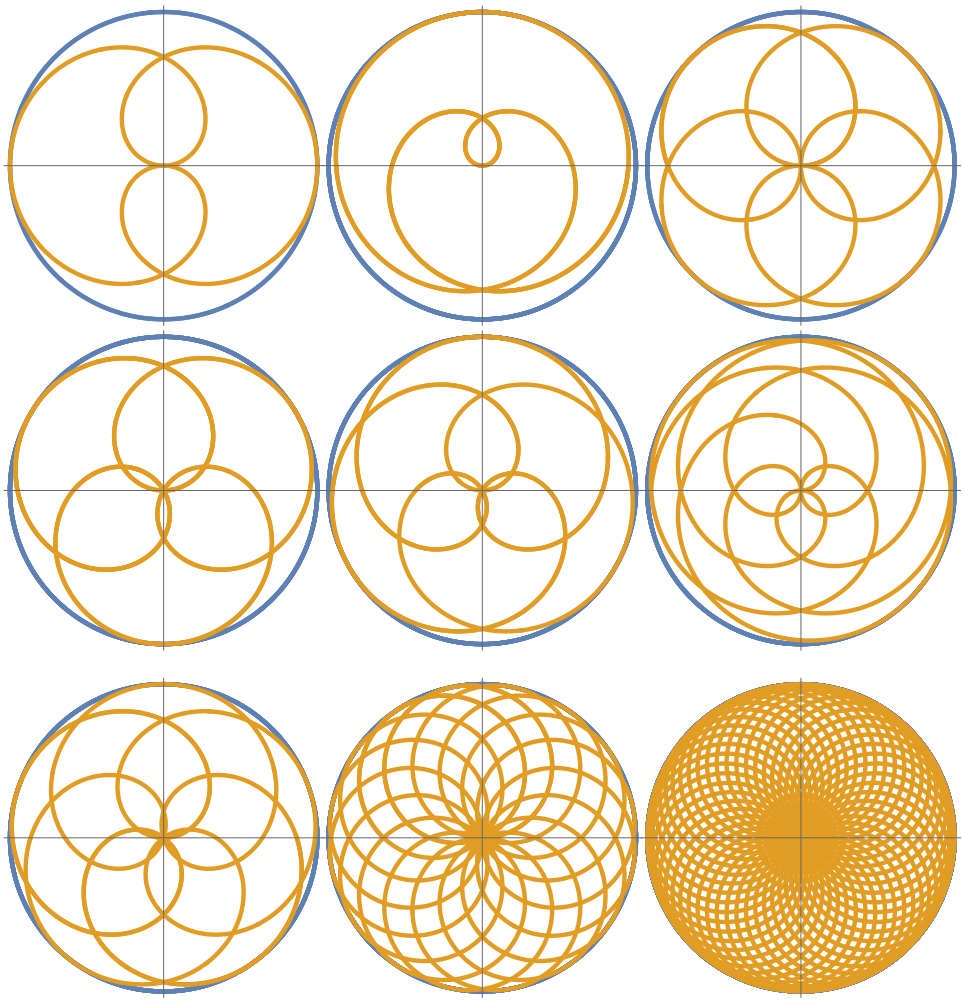
```
r1[t_]:=1+(Abs[Cos[3t]]+2(0.25-Abs[Cos[3t+Pi/2]]))/(2+8Abs[Cos[6t+Pi/2]]);
r2[t_]:=1.8+(Abs[Cos[3t]]+2(0.25-Abs[Cos[3t+Pi/2]]))/(2+8Abs[Cos[6t+Pi/2]]);
PolarPlot[{r1[t],r2[t]}, {t,0,2Pi}, Ticks->False, PlotStyle->{Black,Red}]
```

Figure 1.14: Lotus, by YouTube channel MathsShorts1.




```
par=GoldenRatio;
p=Table[{(n/100)Cos[2par Pi n],(n/100)Sin[2par Pi n]},{n,0,500}];
Graphics[{PointSize[Medium],Red,Point[p]},Axes->True]
par=GoldenRatio^2; (*1/phi^2*)
p=Table[{(Sqrt[n]/2)Cos[2 Pi n/par],(Sqrt[n]/2)Sin[2 Pi n/par]},{n,0,500}];
Graphics[{PointSize[Medium],Red,Point[p]},Axes->True]
par=Pi;
p=Table[{(n/100)Cos[2par Pi n],(n/100)Sin[2par Pi n]},{n,0,500}];
Graphics[{PointSize[Medium],Green,Point[p]},Axes->True]
par=E;
p=Table[{(n/100)Cos[2par Pi n],(n/100)Sin[2par Pi n]},{n,0,500}];
Graphics[{PointSize[Medium],Blue,Point[p]},Axes->True]
```

Figure 1.12: Rosettes based on ϕ, π, and e.



$r = \sin(n\ \theta)$ for various values of n

```
r = 1; n = 24/25;
PolarPlot[{r, Sin[n t]}, {t, 0, 50 Pi}, Ticks -> False]
```

Figure 1.13: Rosettes in Polar Coordinates From $r = \sin(n\theta)$ For Various n.

1.1.7 Egg Shapes

We are familiar with the oval shape of an egg. (The word oval is derived from the Latin "ovus" for egg.) Eggs are mentioned in the Bible and in Shakespeare, and appear in classical painting (Figure 1.17 is from an altarpiece by Piero della Francesca). An egg-like oval is a generalization of the familiar ellipse, much as an ellipse is a generalization of the circle, but how can we derive a mathematical expression for an egg?

We start with the ellipse. The ellipse is defined as the locus of all the points the sum of whose distances from two fixed points, the foci (pronounced fow-sai) of the ellipse, is constant. Figure 1.15a illustrates how this definition leads to a simple construction of the ellipse. A nail is driven into each focus and a string is attached loosely to the nails. A pen is used to tighten the string, and half of the ellipse is drawn by sliding the pen along the string. The string is then moved to the other side of the nails and other half of the ellipse is drawn in the same way.

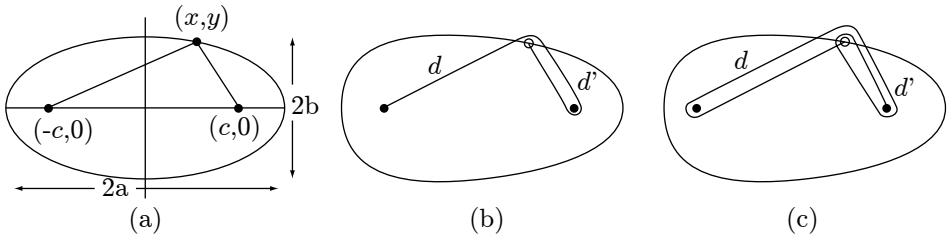


Figure 1.15: Ellipse and Oval.

◇ **Exercise 1.4:** Explain how to succeed in billiard (pool table) with the aid of an ellipse.

The shape of an egg isn't well defined, but the great 19th-century physicist James Maxwell discovered, at the tender age of 15, how to create an entire family of oval shapes that resemble eggs [May 62]. Figure 1.15b shows how to attach a string to the left focus and to the pen (the hollow circle) such that the sum of distance d of the pen from the left focus and twice its distance d' from the right focus is constant. Moving the pen traces an oval shape reminiscent of an egg. Figure 1.16 is a contour plot of the six shapes created by keeping the sum $d+2d'$ equal to 2.5, 3, 3.5, 4, 4.5, and 5. Figure 1.15c shows a similar oval created by attaching a string to the right focus and to the pen in a way that keeps the sum $2d+3d'$ constant. An entire family of oval shapes can be created by keeping the sum $ad+bd'$ constant for various values of a and b.

Here is another approach to the construction of ovals. We modify the equation of the ellipse to give it a left-right asymmetry. The implicit equation of the circle is $x^2+y^2=R^2$ or $(x/R)^2+(y/R)^2=1$ and that of the ellipse is $(x/a)^2+(y/b)^2=1$, where a and b are the two main half-axes of the ellipse. We therefore modify this equation such that the term $(y/b)^2$ will become larger on one side of the x axis and smaller on the other side. The simplest way of doing this is to multiply this term by a simple perturbation function $g(x)$ that satisfies

$$g(x)=\begin{cases}<1 & \text{positive } x,\\ >1 & \text{negative } x.\end{cases}$$

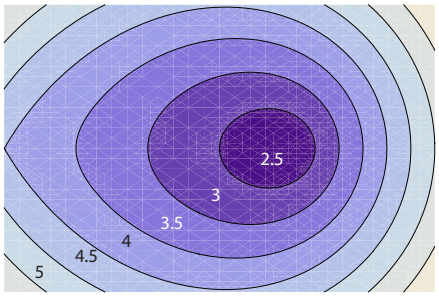
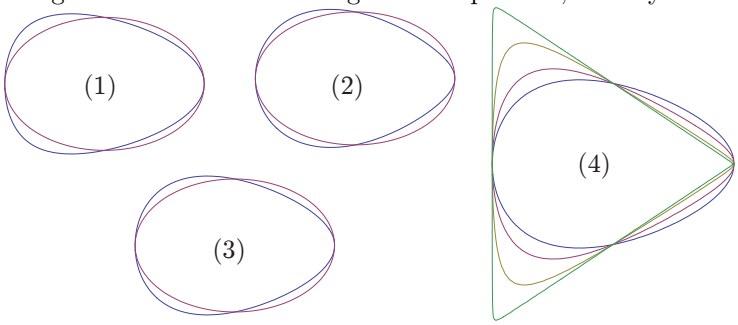


Figure 1.16: Contour Plots of Maxwell Ovals.

As an experiment, we try the implicit expression $(x/a)^2+g(x)(y/b)^2 = 1$ for the functions $g_1(x) = 1 + 0.2x$, $g_2(x) = 1/(1 - 0.2x)$, $g_3(x) = e^{0.2x}$, and $g_4(x) = (1 + kx)/(1 - kx)$. The last perturbation function is computed for $k = 0.1$, 0.2, 0.3, and 0.333 and results in shapes varying from an egg to a triangle. All four functions extend the range of values of y when x is negative and shrink this range when x positive, thereby resulting in ovals.



Section 3.9 discusses a different approach, circle inversions, to constructing and converting curves. Among other features, the circle inversion of an ellipse is an oval.

◇ **Exercise 1.5:** Apply the methods shown here to compute and draw an egg shape in three dimensions.

Figure 1.17 is one of many examples of the use of an egg shape in art and architecture.

1.1.8 The Cassini Oval

The ellipse is the locus of points whose sum of distances from two fixed foci (pronounced fow-sai) is constant. Similarly, the Cassini oval, named after the astronomer Giovanni Cassini, is the locus of points the product of whose distances from two fixed foci is constant.

We denote by b the square root of the constant product of the distances. If the foci are at $(a, 0)$ and $(-a, 0)$, then the equation of the curve is

$$((x - a)^2 + y^2)((x + a)^2 + y^2) = b^4, \quad \text{or} \quad (x^2 + y^2)^2 - 2a^2(x^2 - y^2) + a^2 = b^4.$$

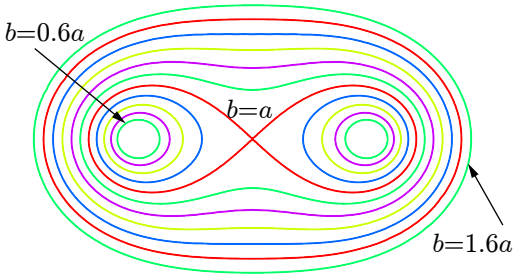
Figure 1.18 shows 11 shapes of this oval, for $a = 1$ and for b values ranging from



Figure 1.17: An Egg by Piero Della Francesca (1465).

$0.6a$ to $1.6a$. For $b/a < 1$, the curve consists of two disconnected loops, each of which contains a focus. For $b/a > 1$, the curve is a single, connected loop enclosing both foci.

The special shape for $b = a$ is the lemniscate of Bernoulli. The name is derived from lemniscatus, Latin for "decorated with hanging ribbons." The lemniscate was first described in 1694 by Jakob Bernoulli as a modification of the ellipse. The central point of the lemniscate is a crunode (or an ordinary double point), a point where a curve intersects itself.



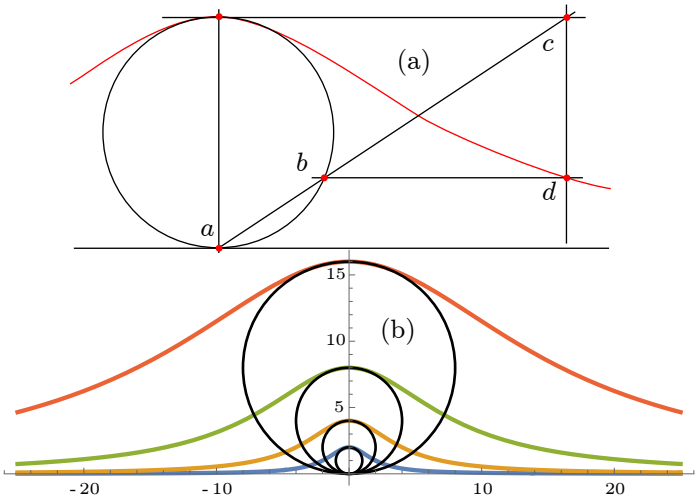

```
a=1;
cassini[x_,y_]:=(x^2+y^2)^2-2a^2(x^2-y^2)+a^4-b^4;
Show[Table[ContourPlot[cassini[x,y]==0, {x,-2,2}, {y,-1.3,1.3},
ColorFunction->Function[{x,y,b},Hue[4b]]], {b,0.6a,1.6a,0.1a}],
Frame->False, AspectRatio->0.6]
```

Figure 1.18: Cassini Ovals for Several b Values.

1.1.9 Witch of Agnesi

The witch of Agnesi is a plane curve that is defined by a generating circle. The Wikipedia article on this curve explains how it is named after Maria Gaetana Agnesi, and how the word "witch," is a result of a meaning lost in translation. Reference [agnesi 24] has much information about this curve.

Figure 1.19a illustrates the special construction of this curve. Draw a line segment from the bottom a of the circle through an arbitrary point b on the circle. Continue the segment until it reaches the height of the circle at c, and then drop a vertical to the height of b. This defines the location of point d, the point on the curve that corresponds to b.




```
Agn[a_]:=8a^3/(x^2+4a^2);
p1=Plot[{Agn[1],Agn[2],Agn[4],Agn[8]}, {x,-25,25}, AspectRatio->Automatic];
p2=Graphics[Evaluate[Table[Circle[{0,a},a],{a,{1,2,4,8}}]]];
Show[p1,p2]
```

Figure 1.19: Construction of the Witch of Agnesi.

Figure 1.19b may help the reader to derive the parametric equation of the curve which is $x = 2a\cot\theta$ and $y = a[1 - \cos(2\theta)]$ for $\theta \in [0, \pi]$. An alternative representation is $x = 2at$ and $y = 2a/(1 + t^2)$ for $t \in (-\infty, \infty)$. A derivation which eliminates t produces the Cartesian equation

$$y = \frac{8a^3}{x^2 + 4a^2}.$$

1.1.10 The Superellipse

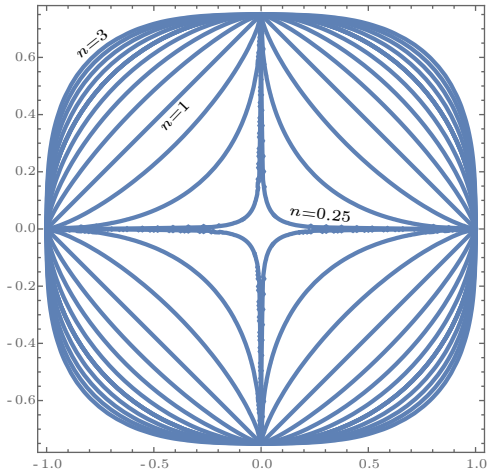
Mathematicians, being curious, always try to extend and generalize current concepts and theories, which is why Gabriel Lamé, a little-known mathematician, decided to extend

the ellipse, a common curve which is itself an extension of the circle. He came up with the simple but flexible expression

$$\left|\frac{x}{a}\right|^n + \left|\frac{y}{b}\right|^n = 1,$$

where a and b, the semi-axes of the superellipse, are positive numbers, and n is a positive real parameter that controls the shape of the curve. For $n = 1$, the superellipse is a rhombus with corners at $(\pm a, 0)$ and $(0, \pm b)$. For $n = 2$, the superellipse is the original ellipse. For $n > 2$, the curve becomes a rectangle with rounded corners, and for $0 < n < 2$ the curve resembles a rhombus with the same corners, but with convex sides.

Figure 1.20 shows several examples for n values form 0.25 to 3, in steps of 0.25.




```
a=1; b=0.75; (*SuperEllipse*)
l={};
SupEllipse[x_,y_]:=(x/a)^n+(y/b)^n-1;
Do[
p1=ContourPlot[SupEllipse[x,y]==0, {x,-a,a}, {y,-b,b}];
p2=ContourPlot[SupEllipse[-x,-y]==0, {x,-a,a}, {y,-b,b}];
p3=ContourPlot[SupEllipse[x,-y]==0, {x,-a,a}, {y,-b,b}];
p4=ContourPlot[SupEllipse[-x,y]==0, {x,-a,a}, {y,-b,b}];
p5=Show[{p1,p2,p3,p4},ColorFunction->Function[{n},Hue[4n]]];
AppendTo[l,p5], {n,0.25,3,0.25}]
Show[l]
```

Figure 1.20: SuperEllipses for Several n Values.

Because of the use of a real parameter, the superellipse can take on many shapes and has been used over the years to design typefaces, a map projection (the Tobler hyperelliptical projection), various commercial logos, the 1968 Azteca Olympic Stadium in Mexico City, and the shapes of app icons in versions 7 and up of the popular iOS operating system. Perhaps the most well-known application of the superellipse is Sergels

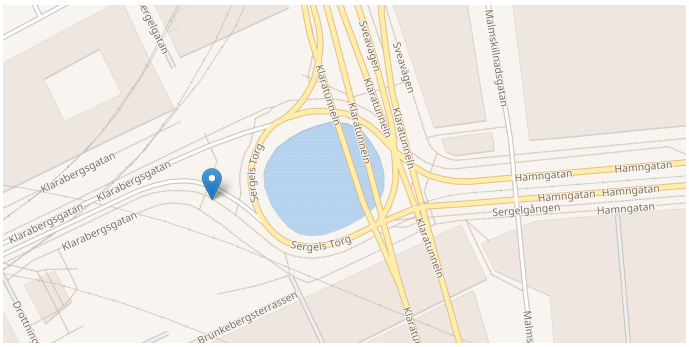


Figure 1.21: Sergels Torg, Stockholm.

torg, a public square in Stockholm, Sweden (Figure 1.21), designed by Piet Hein in the 1960s.

> There is
> one art,
> no more,
> no less:
> to do
> all things
> with art-
> lessness.
> —Piet Hein, *Grooks*.

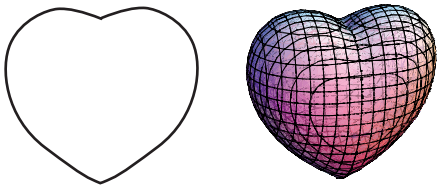
Piet Hein was also responsible for the solid object known as a superegg. This is the surface of revolution obtained by rotating a superellipse (with an exponent greater than 2) about its longer axis. The curvature of this solid at its top and bottom is zero, which allows it to stand upright on a flat surface

1.1.11 Heart Shape

The classical shape of a heart is generally considered attractive. Two-dimensional heart curves are easy to generate by (experimentally derived) formulas such as $(x^2+y^2-1)^3-x^2y^3=0$, $x^2+2(\frac{3}{5}x^{2/3}-y)^2-1=0$, and $r=5\sin^7(t)e^{|2t|}$ for $-\pi\le t\le\pi/2$ and $\pi/2\le t\le\pi$. In contrast, a three-dimensional "heart" surface is much more difficult to express mathematically because it is smooth all over except for sharp point at the bottom and a deep indentation at the top. The heart shown in the figure follows the implicit equation

$$(x^2+(1.5y)^2+z^2-1)^3-x^2z^3-(1.5y)^2z^3/20=0,$$

that was derived experimentally by Gabriel Taubin of IBM [Taubin 94].



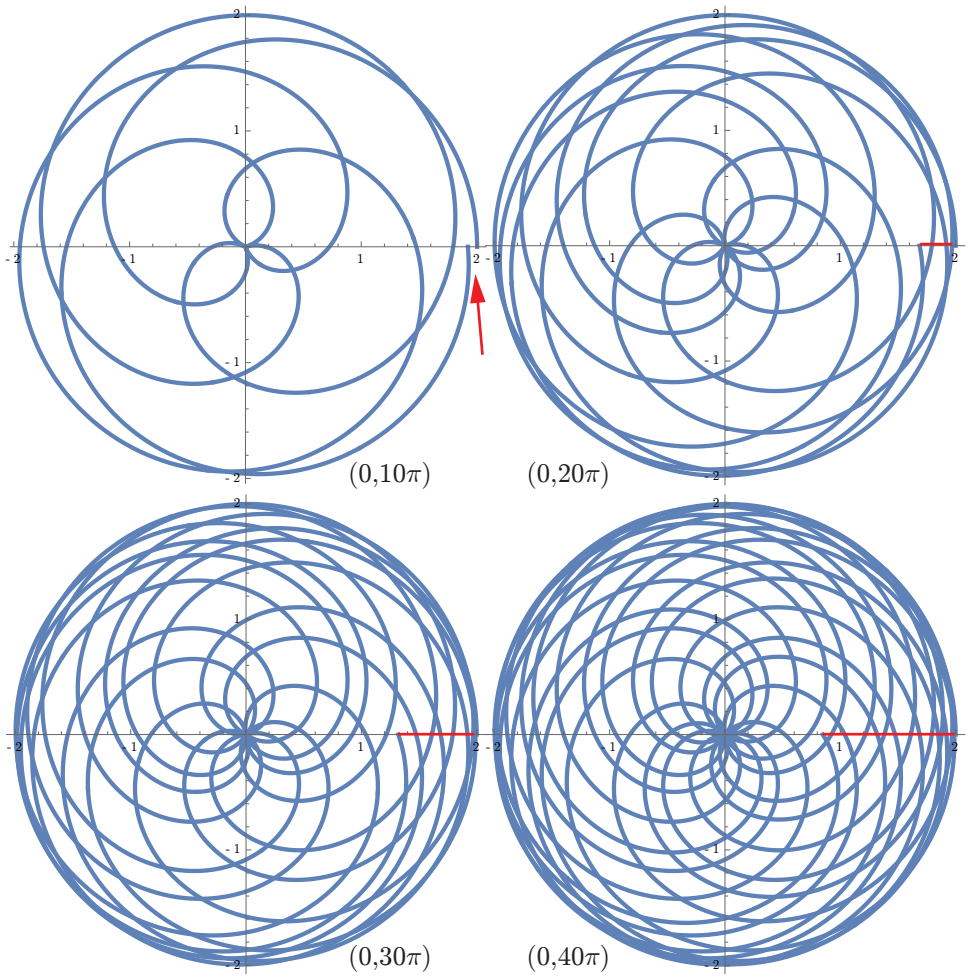
1.1.12 Irrationality in a Periodic Function

Figure 1.22 shows the effect of having an irrational value (φ) as part of an otherwise periodic function. The function is $r = e^{i\theta} + e^{i\theta\varphi}$ and the four parts of the figure are its plots for θ from 0 to 10π, 20π, 30π, and 40π. Having any rational value instead of φ would cause this function to be periodic, but the red segments indicate how the irrational value of φ causes the plot to deviate more and more from true periodicity.

1.2 Perspective

Painters, engineers, and architects have recognized in the early 1400s that paintings and drawings of three-dimensional objects look realistic if drawn following two simple rules: (1) closer objects should appear small and (2) groups of parallel lines on the object should appear to converge to a vanishing point (except if the lines are perpendicular to the line of sight). This technique became known as linear perspective (Section 8.23), and it is an important concept in art, and technical drawing. Using a computer, it is quick and easy to perform the vast amount of computations needed and display figures in perspective. The figure shown here depicts two train tracks and it is obvious that the lines, which are parallel, converge to the vanishing point at the center of the figure.






```
PolarPlot[{Abs[E^(I t)+E^(GoldenRatio I t)],Arg[t]},{t,0,10 Pi}]
```

Figure 1.22: Polar Plot Involving an Irrational Value.

1.3 Moiré Patterns

A moiré pattern is a large-scale interference pattern created by overlapping small-scale patterns of lines or other shapes. According to Wikipedia, the term originates from moire (moiré in its French adjectival form), a type of textile, traditionally made of silk but now also made of cotton or synthetic fiber, with a rippled or "watered" appearance. Moiré patterns have important applications in marine navigation (as special lights that point to hazards or to safe passages), to protect banknotes from forgery, and in high-resolution microscopy.

The Wikipedia article on moiré discusses several ways to create such a pattern, and

more are shown in reference [Moire 25]. The first four patterns of Figure 1.23, based on the entertaining video at [Moire 15], start with a pattern of 3500 random dots, generated by the Mathematica code:

```
lc = {}; (*Random points for a Moire pattern*)
Do[AppendTo[lc, Point[{Random[Real, 1.3], Random[]}]], {i, 3500}];
Graphics[lc]
```

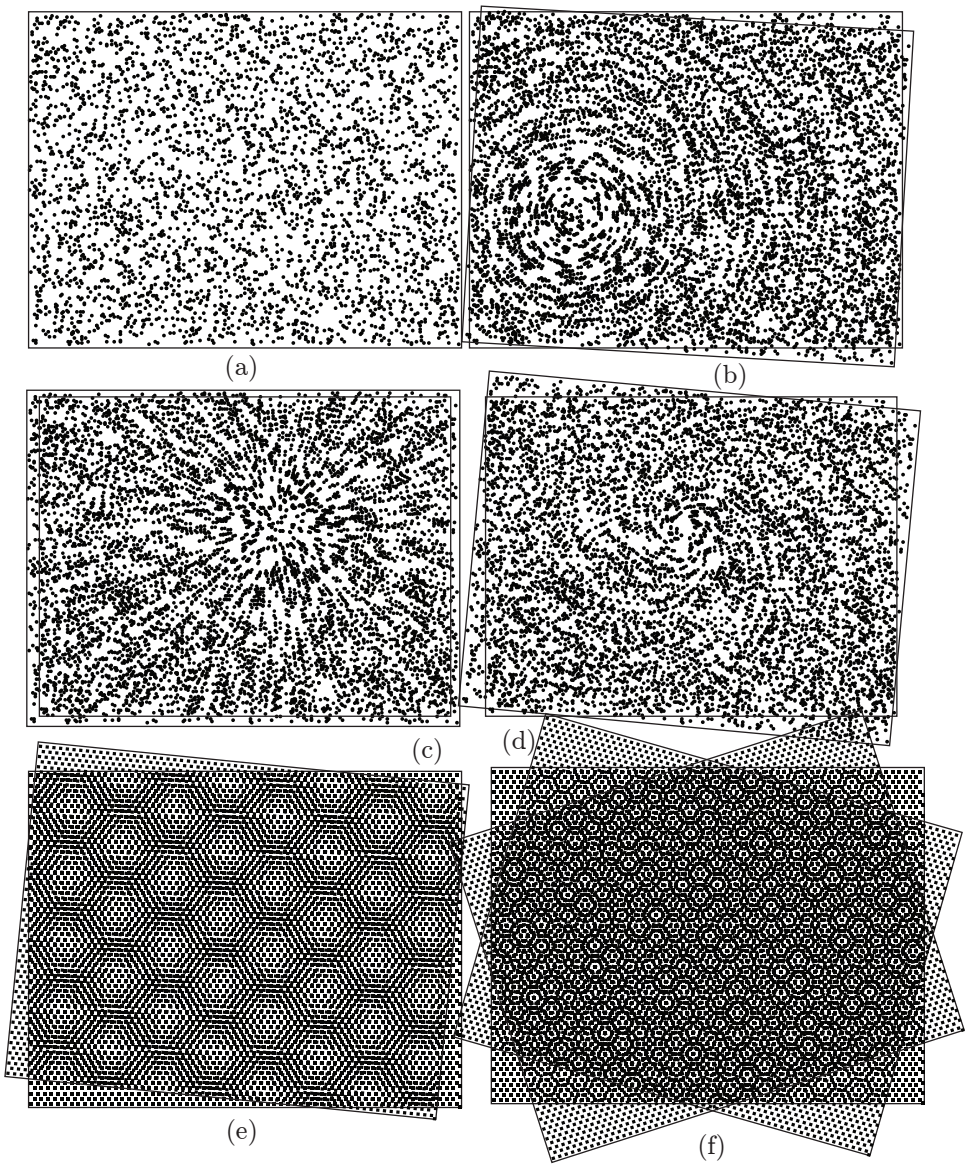


Figure 1.23: Moiré Patterns.

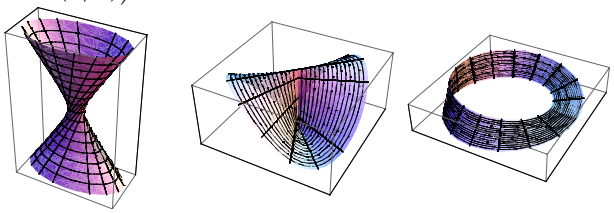
Part (b) of the figure was obtained when a copy of the original was slightly shifted and rotated. The circular patterns are a product of our minds, the result of the fact that points close to the center of rotation are moved only slightly by the rotation, while points away from the center are moved more. The starburst of part (c) is the result of shrinking the original by 95%. The spirals of part (d) are due to a small rotation of the copy.

These moirés are a consequence of the Brouwer fixed-point theorem, Section 10.10. This theorem guarantees that when a smaller, transparent copy of an image is placed over the image and is translated or rotated while kept fully inside the original image, there would always be an image point in the copy that is placed directly over its original.

Parts (e) and (f) of the figure are simple moirés created by a rectangular dot pattern.

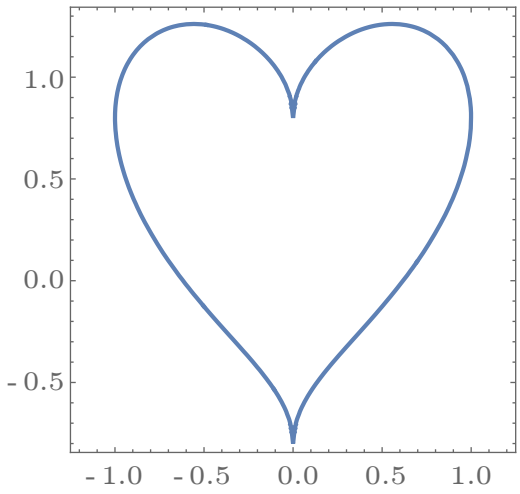
1.4 Ruled Surfaces

A ruled surface is generated by connecting two space curves with straight segments. Given two space curves, $P(u)$ and $Q(u)$, a ruled surface is constructed between them by drawing a family of straight segments that connect points on $P(u)$ to points on $Q(u)$. Given two points, a and b, the equation of the straight segment connecting them is $(1-w)a + w\,b$ for $0 \le w \le 1$, which is why the parametric representation of a ruled surface is $(1-w)P(u) + wQ(u)$ where the two parameters u and w vary independently. The figure shows three ruled surfaces. The one-sheet hyperboloid is the sum $(1-w)\big(a\cos u, b\sin u, 0\big) + w\big(-a\sin u, b\cos u, c\big)$, the cylindroid (also known as Plücker's conoid) is given by the sum $(1-w)\big(0, 0, 2\cos u \sin u\big) + w\big(\cos u, \sin u, 0\big)$, and the well-known Möbius strip of radius a and width $2w$ is obtained by $\big((a+w\cos(u/2))\cos(u), (a+w\cos(u/2))\sin(u), w\sin(u/2)\big)$.



1.5 Most Important Curve

And finally, the most important curve of all (whose source I have unfortunately forgotten).




```
ContourPlot[x^2+(5y/4-Sqrt[Abs[x]])^2==1,{x,-1.2,1.2},{y,-0.8,1.3}]
```

Figure 1.24: Most Important Curve.

1.6 Listings of Mathematica codes

```
g1=ParametricPlot3D[{Sin[3t],Cos[3t],t/3}, {t,0,2Pi}, Ticks->None]
g2=Plot3D[Cos[x y], {x,0,Pi}, {y,0,Pi}, Ticks->None];
g3=Plot3D[Exp[-(x^2+y^2)], {x,-2,2}, {y,-2,2}, Ticks->None]
Show[GraphicsArray[{g1,g2,g3}], ImageSize->72 6]
```

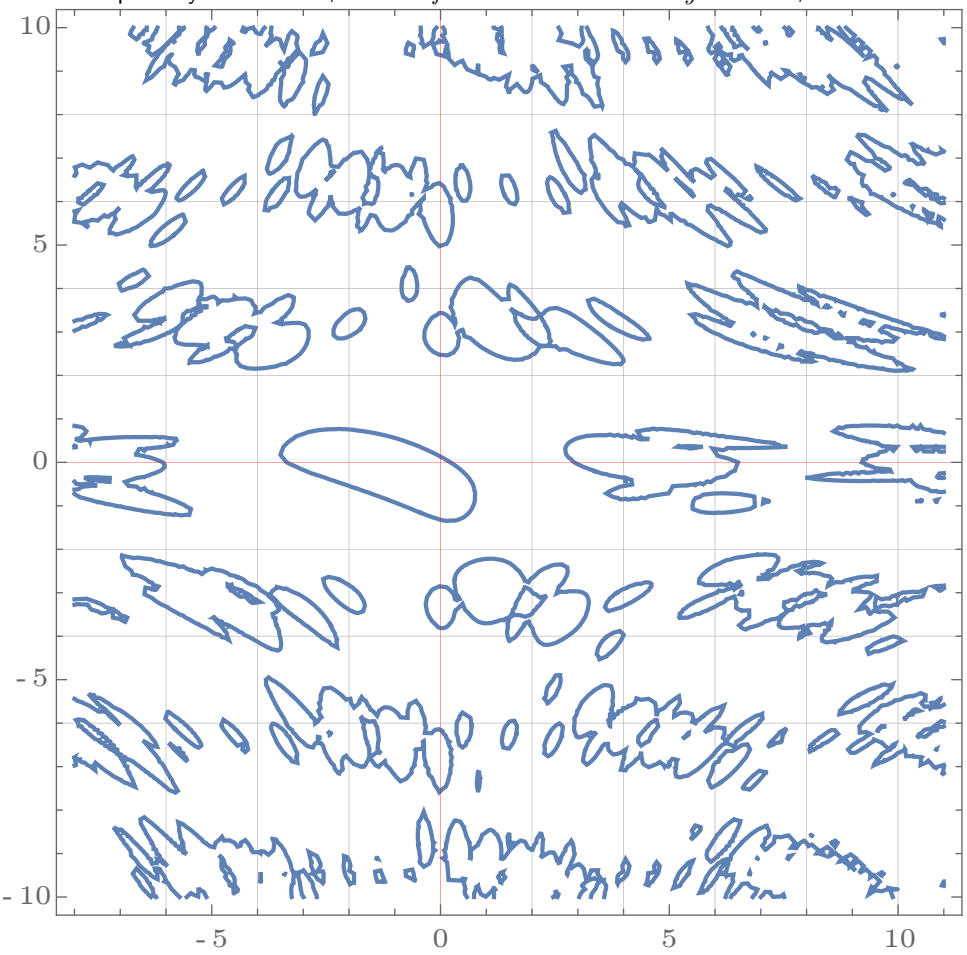
Three-Dimensional Plots.

```
$TextStyle={FontFamily->"CMR10", FontSize->10}
g1=Show[Graphics[{AbsolutePointSize[3],{Point[{1,1.25}],Point[{0.4,0.9}],
 Point[{-1,-0.5}]}}], Axes->True, AxesLabel->{"x","y"}, AspectRatio->1];
g2=Plot[x^4, {x,1,2}];
g3=Plot[Sin[x], {x,0,Pi}];
g4=Show[Graphics[Line[Table[{n,Sin[n]},{n,0,Pi,0.75}]]], Axes->True];
Show[GraphicsArray[{{g1,g2},{g3,g4}}], ImageSize->72 6]
```

Points and Curves.

```
(* GRParSurfC parametric surfaces in color *)
g1=ParametricPlot3D[{w Cos[u],w Sin[u],(w Cos[7u]/2)/2,
  Hue[u/6.5]}, {u,0,2Pi}, {w,0,1}, Axes->False,
  Boxed->False, PlotPoints->{55,10},
  Mesh->None]; (* flower *)
g2=ParametricPlot3D[{w Cos[u],w Sin[u],0.6u}, {u,0,4Pi}, {w,0,2.5},
 Axes->False, Boxed->False, PlotPoints->{25,9},
 ViewPoint->{2,-4,4}, PlotStyle->Directive[RGBColor[0,.8,.7],
 Specularity[1, 5]]];(* helicoid *)
```

Inspired by Julian Havil, *Curves for the Mathematically Curious*, 2019




```
ContourPlot[Sin[Cos[2y]-Sin[x+y]]==Cos[Sin[x+2y]-Cos[x y]],
 {x,-8,11},{y,-10,10},
 GridLines->{{-10,-6,-4,-2,{0,Directive[Red]},2,4,6,8,10},
 {-6,-4,-2,{0,Directive[Red]},2,4,6,8,11}}]
```

Figure 1.25: Mathematical Doodling, $\sin(\sin(x)+\cos(y)) = \cos(\sin(xy)+\cos(x))$.

```
g3=ParametricPlot3D[{w Cos[u],w Sin[u],w^2, RGBColor[.8,0,.7]},
 {u,0,2Pi}, {w,0,0.8}, Axes->False, Boxed->False,
 PlotPoints->{15,9}, Mesh->None]; (* paraboloid *)
g4=ParametricPlot3D[{w Cos[u],w Sin[u],Cos[w]}, {u,0,2Pi}, {w,0,5Pi},
 Axes->False, Boxed->False, PlotPoints->{25,20},
 PlotStyle->Directive[RGBColor[0.8,0,0.2]]]; (* surface of revolution *)
g5=ParametricPlot3D[{w Cos[u],w Sin[u],Cos[w-4u],Hue[u/6]},
 {u,0,2Pi}, {w,0,5Pi}, Axes->False, Boxed->False,
 PlotPoints->{30,15}, Mesh->None]; (* a variation *)
Show[GraphicsRow[{g1, g2, g3, g4, g5}], ImageSize->72 5]
```

Code for Figure 1.3

```
Hypot[a_, b_]:=ParametricPlot[{(a - b) Cos[t] +
  h Cos[(a - b) t/b], (a - b) Sin[t] - h Sin[(a - b) t/b]},
 {t, 0, a Pi}, Axes -> False];
GraphicsRow[Table[Hypot[3, 1], {h, 0.5, 3.5, 0.5}]]
GraphicsRow[Table[Hypot[4, 1], {h, 0.5, 3.5, 0.5}]]
GraphicsRow[Table[Hypot[5, 1], {h, 0.5, 3.5, 0.5}]]
GraphicsRow[Table[Hypot[5, 2], {h, 0.5, 3.5, 0.5}]]
GraphicsRow[Table[Hypot[7, 2], {h, 0.5, 3.5, 0.5}]]
GraphicsRow[Table[Hypot[7, 3], {h, 0.5, 3.5, 0.5}]]
```

Code for Figure 1.4.

```
k = 25;
Manipulate[
 Show[Graphics[Table[{Hue[n/25],
 Line[{{Cos[n a], Sin[n a]}, {Cos[(n + 1) a], Sin[(n + 1) a]}}]},
 {n, 0, k}]]],
{a, 0, Pi, 0.1}]
```

Seven polyline patterns in a circle.

```
k = 45;
Manipulate[Show[Graphics[
 Table[{Line[{{4Cos[n a]+Cos[4n a],
 4Sin[n a]-Sin[4n a]},{4Cos[(n+1)a]+Cos[4(n+1)a],
 4Sin[(n+1)a]-Sin[4(n+1)a]}}]}, {n, 0, k}]]],
 {a, 0, 5Pi, 0.1}]
```

Six polyline patterns in a five-sided hypotrochoid.

```
tile0[a_,b_]:=Union[{Circle[{a+1,b},0.5, {Pi/2,Pi}],
  Circle[{a,b+1},0.5, {3Pi/2,2Pi}]}];
tile1[a_,b_]:=Union[{Circle[{a,b},0.5, {0,Pi/2}],
  Circle[{a+1,b+1},0.5, {Pi,3Pi/2}]}];
Show[Graphics[
 Table[Subscript[tile, Random[Integer]][a,b],{a,1,10},{b,1,10}]],
 AspectRatio->Automatic]
```

Two patterns of circular arcs randomly selected and packed in a 10×10 grid.

```
tile0[a_,b_]:=
  Union[{{Hue[a/10], Disk[{a+1,b},0.5, {Pi/2,Pi}]}, {Hue[b/10],
  Disk[{a,b+1},0.5, {3Pi/2,2Pi}]}}];
tile1[a_,b_]:=
  Union[{{Hue[b/10], Disk[{a,b},0.5, {0,Pi/2}]}, {Hue[a/10],
  Disk[{a+1,b+1},0.5, {Pi,3Pi/2}]}}];
Show[Graphics[
  Table[Subscript[sile, Random[Integer]][a,b], {a,1,10}, {b,1,10}]],
 AspectRatio->Automatic]
```

Two patterns of circular disks randomly selected and packed in a 10×10 grid.

```
PolarPlot[{t, E^0.1 t}, {t,
  0, 12 Pi}, Ticks -> False,
 PlotStyle -> {Black, Directive[Dashed, Thick, Red]}]
(* +-sqrt[t] spirals *)
PolarPlot[{Sqrt[t], -Sqrt[t]}, {t, 0, 12 Pi},
 Ticks -> False, PlotStyle -> {Black, Directive[Dashed, Red]}]
```

Archimedean and logarithmic spirals.

```
tile[a_,b_]:=Union[{Disk[{a+1,b},0.5, {Pi/2,Pi}],
  Polygon[{{a,b}, {a,b+1}, {a+1,b+1}}]}];
 Show[Graphics[Table[{Hue[RandomReal[]], Rotate[ tile[m, n],
 90 Random[Integer,{0,3}] Degree, {m+1/2,n+1/2}]},
 {m,1,10}, {n,1,10}]]]
```

A square pattern rotated and packed to form a 10×10 grid for exercise.

```
a = 3; b = 2; g = E^(0.2 x);
ContourPlot[{(x/a)^2+g(y/b)^2==1, (x/a)^2+(y/b)^2==1},
 {x, -a, a}, {y, -b-0.2, b+0.2}, Frame->False,
 AspectRatio -> Automatic]
```

Code for Eggs generated by perturbing an ellipse

```
a = 3; b = 2; g[k_]:=(1+kx)/(1-kx);
ContourPlot[{(x/a)^2 + g[0.1] (y/b)^2 ==1,
 (x/a)^2+g[0.2](y/b)^2==1, (x/a)^2+g[0.3](y/b)^2==1,
 (x/a)^2+g[0.333](y/b)^2==1}, {x,-a,a}, {y, -2b, 2b},
 Frame -> False, AspectRatio -> Automatic]
```

Code for eggs generated for four values of parameter k.

```
a=1; b=1; c=2; g=1+0.25z;
ContourPlot3D[g(x/a)^2+g(y/b)^2+(z/c)^2==1, {x,-a-0.2,a+0.2},
 {y,-b-0.2, b+0.2}, {z,-c,c}, Axes->False, BoxRatios->Automatic]
```

Code for generating a three-dimensional egg shape by perturbing an ellipsoid

```
ContourPlot3D[(x^2+(1.5y)^2+z^2-1)^3-x^2 z^3-(1.5y)^2 z^3/20==0,
 {x,-1.3,1.3}, {y,-0.8,0.8}, {z,-1,1.3}, Axes->False,
 BoxRatios->Automatic, Mesh->None, ColorFunction->(Hue[#3] &)]
```

A heart shape (after Gabriel Taubin).

```
a=2; b=1; c=3;
g1=ParametricPlot3D[{a(Cos[u]-w Sin[u]),b(Sin[u]+w Cos[u]),cw},
 {u,0,2Pi}, {w,-2Pi,2Pi}, Ticks->None];
g2=ParametricPlot3D[{w Cos[u],w Sin[u],2Cos[u]Sin[u]}, {u,0,2Pi},
 {w, 0, 2}, Ticks->None];
g3=ParametricPlot3D[{(a+w Cos[u/2])Cos[u],
 (a+w Cos[u/2])Sin[u],w Sin[u/2]}, {u, 0, 2Pi}, {w, -0.5, 0.5},
   Ticks->None];
Show[GraphicsRow[{g1, g2, g3}]]
```

Ruled surfaces, Hyperboloid, Cylindroid, and Möbius strip.

I'm a creative guy, artistically with graphics.
—Tim Allen.



2
Numbers: The Basic Building Blocks

Numbers play a fundamental role in forming various mathematical entities and structures. The average person discusses numbers as if there is only one type, but mathematicians distinguish many types of numbers, the most important of which are the integers, rationals, irrationals, transcendental, reals, imaginary, and complex. These are discussed in this chapter. Reference [Conway 96], for advanced readers, discusses other types of numbers such as quadratic surds, quaternions, octonions, Cayley numbers, transfinite numbers, and infinitesimal numbers.

2.1 Arithmetic Operations

We are familiar with the basic arithmetic operations such as addition and subtraction, but mathematicians, being methodical, always try to generalize concepts, which is why they came up with an infinite sequence of hyperoperations which are denoted by $H_n(a, b)$ or by $a[n]b$. The first few hyperoperations are summarized in Table 2.1.

> "Reeling and Writhing of course, to begin with," the Mock Turtle replied, "and the different branches of arithmetic—ambition, distraction, uglification, and derision."
>
> —Lewis Carroll, *Alice in Wonderland*, 1865.

The name of the nth member of the hyperoperation sequence starts with the prefix of the Greek name of n, followed by the string "`ation`." Thus, we have tetration for $n = 4$, pentation for $n = 5$, hexation for $n = 6$, and so on. The following is a recursive

definition of the hyperoperations for $n \geq 2$:

$$H_n(a,b) = a[n]b = \begin{cases} b+1 & \text{if } n = 0; \\ a & \text{if } n = 1 \text{ and } b = 0; \\ 0 & \text{if } n = 2 \text{ and } b = 0; \\ 1 & \text{if } n \geq 3 \text{ and } b = 0; \\ H_{n-1}(a, H_n(a, b-1)) & \text{otherwise.} \end{cases}$$

Name	Rank	Definition
Succession	0	$1+b$
Addition	1	$a+b$
Multiplication	2	$a \times b$
Exponentiation	3	a^b
Tetration	4	$\underbrace{a^{a^{\cdot^{\cdot}}}}_{b}$

Table 2.1: The Hyperoperation Sequence of Arithmetic Operations.

A related topic is the representation of large numbers. Back in 1928, Wilhelm Ackermann came up with a compact notation that made it easy to write extremely large numbers, numbers for which the conventional scientific (or exponential) notation requires much space and becomes cumbersome and unreadable. Ackermann's notation was improved and popularized by Donald Knuth in 1976 and is now known as the up-arrow notation. See also Page 66.

For $n = 3$, the conventional notation a^b becomes $a \uparrow b$, and higher-rank hyperoperations are similarly denoted with more and more up-arrows according to:

$$\begin{aligned} H_4(a,b) &= a \uparrow\uparrow b, \\ H_5(a,b) &= a \uparrow\uparrow\uparrow b, \\ &\vdots \\ H_n(a,b) &= a \uparrow^{n-2} b, \text{for } n \geq 3. \end{aligned}$$

The pair $(2,4)$ is used here to illustrate the power and compactness of this notation. We start with a single arrow (exponentiation)

$$2 \uparrow 4 = H_3(2,4) = 2 \times (2 \times (2 \times 2)) = 2^4 = 16.$$

Continue with a double arrow (tetration)

$$2 \uparrow\uparrow 4 = H_4(2,4) = 2 \uparrow (2 \uparrow (2 \uparrow 2)) = 2^{2^{2^2}} = 2^{16} = 65{,}536.$$

Only one more step is needed to convince anyone that the up-arrow notation is ideally suitable for the representation of extremely large numbers. This step is based on the

general rule is that an n-arrow hyperoperator expands into a right-associative series of $(n-1)$-arrow hyperoperators. Thus,

$$2 \uparrow\uparrow\uparrow 4 = H_5(2,4) = 2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 2)) = \underbrace{2 \times (2 \times (2 \times \ldots))}_{2\uparrow\uparrow 4 \text{ copies of } 2} = \underbrace{2^{2^{\cdot^{\cdot^{2}}}}}_{65{,}536\ 2\text{'s}} .$$

If number be lacking
It maketh men dumb
So that to most questions
They must answer mum.
—Robert Recorde.

2.2 Logical Operations

The integer 2 is special as the base of a numbering system. We use decimal numbers (base 10) because we have 10 fingers, but other numbers, both integers and non-integers, can serve as a base. Base 2 is special because it offers an extra set of operations, the logical operations.

Numbers are useful because we can operate on them. Without operations, numbers, whether integers or real, would have limited use. In addition to the familiar arithmetic operations, numbers to any base can also be shifted. Shifting a decimal number one position to the left amounts to multiplying it by 10, while shifting it one position to the right amounts to dividing it by 10, and possibly losing its least-significant digit.

However, binary (base 2) numbers are different in this respect, and are also very useful, because they offer the important family of logical operations. One of those operations, negation, inputs one bit and outputs its opposite value. The other logical operations input two bits each and output one bit. The definitions of the most important logical operations are listed in Table 2.2 as truth tables, but there is a total of 16 such operations, because a binary truth table consists of four bits, which can have $2^4 = 16$ values.

	A	1	1	0	0
	B	1	0	1	0
negation	¬A	0		1	
conjunction (AND)	A∧B	1	0	0	0
disjunction (OR)	A∨B	1	1	1	0
exclusive OR (XOR)	A⊕B	0	1	1	0
equivalence	A≡B	1	0	0	1

Table 2.2: Several Logical Operations.

The logical operations are important because (1) they are easy to implement in hardware, because it is much simpler to design and make electronic circuits that have two states than 10 or any other number of states, and (2) it is possible to implement the all-important arithmetic operations on binary numbers by means of logical operations. This is one reason why digital computers use binary numbers.

◇ **Exercise 2.1:** Why were most early computers decimal?

Pythagoras says that number is the origin of all things; certainly, the law of numbers is the key that unlocks the secret of the universe [Carus 08].

A number is an abstract way to represent a quantity.

A number is a mathematical object (or a concept or an arithmetic value) used to count, measure, and label.

An arithmetical value, expressed by a word, symbol, or figure, representing a particular quantity and used in counting and making calculations and for showing order in a series or for identification.

2.3 Integers

We start with the simplest numbers, the integers. We are familiar with the integers and have been using them from an early age. Yet mathematicians keep discovering properties of these numbers, relations between them, and surprising and unexpected facts about them; an activity that never ends. The study of the integers is known as number theory and it keeps many researchers awake at night, wondering how to prove theorems about the integers or how to find counter-examples that would disprove certain conjectures.

We first describe a fun experiment with integers, based on reference [IntegerProport 23], that will make us respect the integers and wonder about their infinity. We want to see how many times a certain digit, say, 3, appears in a set of consecutive integers. A direct check reveals that in the interval $[1, 10)$, i.e., positive integers less than 10, only one, namely 3, contains the digit 3, a 10% ratio or coverage. In the interval $[1, 100)$, i.e., integers less than 10^2, there are 19 that contain the digit 3, namely the 10 numbers 3, 13, 23,... up to 93, plus the 10 numbers of the form $3x$, minus 33 (which was counted twice); a proportion of 19%.

In the interval $[1, 1000)$, i.e., integers less than 10^3, a similar count reveals 271 numbers that contain the digit 3, a coverage of 27.1%. Going over the first 1000 positive integers and checking for occurrences of 3, yields the following interesting connection. The value 271 was obtained by multiplying the previous 19 by 9 and adding 100 (or 10^2).

In general, we denote the number of integers in the interval $[1, 10^n)$ that contain a 3 by T_n and can now write $T_{n+1} = 9T_n + 10^n$. When applied to $n = 4$, i.e., integers in the interval $[1, 10^4)$, we get $T_4 = 9 \times 271 + 10^3 = 3439$ or 34.39% coverage.

Before we continue, here is another way to compute these coverages or proportions. Each digit in 4-digit integer can be chosen from the 10 decimal digits. The total number of choices (or integers) is therefore $10^4 = 10000$. We are interested in numbers that

contain a 3, so let's invert our problem and count the number of integers that do not have a 3. In this case, we have four-digit numbers where there are only nine choices for each digit, for a total of $9^4 = 6561$ or $10^4 - 9^4 = 10000 - 6561 = 3439$, very elegant!

In general, we can write $T_{n+1} = 10^n - 9^n$, from which we get the general coverage or proportion, of the digit 3 as

$$\frac{T_{n+1}}{10^n} = \frac{10^n - 9^n}{10^n} = 1 - \left[\frac{9}{10}\right]^n \Rightarrow \lim_{n\to\infty} \frac{T_{n+1}}{10^n} = 1.$$

It seems that more and more integers contain the digit 3, but of course there is nothing special about 3. Our surprising, unexpected result means that as we examine more integers, the proportion of any digit grows toward 100%. This is one of the many curious and baffling aspects of the infinite.

The following fact is a weird result of the previous analysis. The harmonic series diverges, but if we remove every term which has a d in it—where d is any integer, single- or multi-digit—the series would converge to about 22.92.

Next, we examine certain integers and describe some of their interesting, unusual, elegant, and even beautiful properties.

0. Zero is interesting because it provides the basis of the modern mathematical notation. This place-value system uses powers of 10 and also employs zeros as place holders. Thus, the string 4080 means $4{\cdot}10^3{+}0{\cdot}10^2{+}8{\cdot}10^1{+}0{\cdot}10^0 = 4000{+}80$. The value of the digit zero is nothing (null), yet it plays an important role in mathematics because of its use as a place holder. In contrast, the Roman numeral for 4080 is `mmmmlxxx`, which is difficult to write and even harder to operate on.



Figure 2.3: Image by MS Copilot, Sept. 2025.

◇ **Exercise 2.2:** Back in the late 1990's, many people believed that the 21st century would start in year 2000. Currently, more and more people realize that this century started in 2001 and will end in year 2100. It will include years 2001, 2002, ..., 2099, 2100. Similarly, the 18th century started in year 1701 and consisted of years 1701, 1702, ..., 1799, 1800. Such numbering is confusing. It would make more sense if the 18th century included all the 100 years whose numbers start with 18, from 1800 to 1899. Why

is our calendar based on such a confusing system, instead of the simpler 2000–2099 and 1700–1799?

Other reasons why zero is interesting are: It is the additive/subtractive unity; $a \pm 0 = a$ for any number a. The product $a \cdot 0 = 0$ for any number a. Zero is the only number that is neither positive nor negative. A division by zero is undefined. Zero is one of the only two digits in the binary number system, and this system is important because of its use in computers and other digital devices. Zero is also the first even nonnegative integer.

The product $a \times 0$ is zero for any number a because multiplication is repeated addition. Multiplying a by 3 is done by adding three copies of a, which is why multiplying a by 0 is done by adding zero copies of a, thereby resulting in zero. Alternatively, this product can also be interpreted as adding zero to itself five times, which produces the same result.

Dividing by zero is undefined because of the following obvious reason. When a number a is divided by smaller and smaller numbers, the result gets bigger and bigger, leading a naive user to assume that in the limit, dividing $a/0$ produces infinity. This idea leads to wrong results such as

$$a/0 = \infty \text{ and } b/0 = \infty \rightarrow a = b.$$

We therefore agree that infinity, even though an important mathematical and philosophical concept, is not a number and any mathematical expression in which it is used or produced is invalid.

However, the symbol ∞ is valid as an indication of an upper or lower limit in expressions that require limits, such as integration and repeated summation. Thus, for example, the expression $a/0 = \infty$ is invalid, but $\lim_{x\to 0^+}(a/x) = \infty$ is valid.

The concept of zero and the modern positional number system were introduced in Europe by Leonardo of Pisa, also known as Fibonacci.

> Zero is... a fine and wonderful refuge of the divine spirit—almost an amphibian between being and non-being.
>
> —Gottfried Wilhelm Leibniz.

1. The number 1 is interesting because it is the first nonzero integer, the first odd integer, and the multiplicative identity, i.e., $a \cdot 1 = a$ for any number a. This integer is one of the two digits of the important binary system. It equals its factorial, its own square, cube, etc. Because $1^n = 1$, the function 1^n has no unique inverse. This is why 1 cannot serve as the base of logarithms.

This number cannot serve as the base of a number system. Recall that decimal numbers use 10 digits and binary (base 2) numbers use two digits. A number system based on 1 would have just one digit, but a positional number system requires a zero, which is why it needs at least two digits.

The first two Fibonacci numbers are 1 and it is well known (see, for example, [Sloane 08]) that many sequences of integers start with 1. A prime number is divisible by itself and by 1, whereas a composite number is divisible by several numbers. Because of these definitions, the integer 1 is considered neither prime nor composite.

◇ **Exercise 2.3:** Observe that when 1 is added to an even number it produces an odd number and when it is added to an odd number it produces an even number. Are we then justified in considering 1 an even/odd hybrid?

2. The integer 2 is interesting because it is the first even number, the first prime number, the only even prime, and the smallest integer that can serve as a concise base of a positional number system. If n is a concise base of such a system, then the representation of any number a in this system is essentially $(\log_n a)$-digits long. Notice that 1 can be used as the basis of a number system, but it is not concise, because the length of a number a in such a case would be a digits. For base-2, the precise length of the binary representation of a is

$$1 + \lfloor \log_2 a \rfloor = \lceil \log_2 (a + 1) \rceil.$$

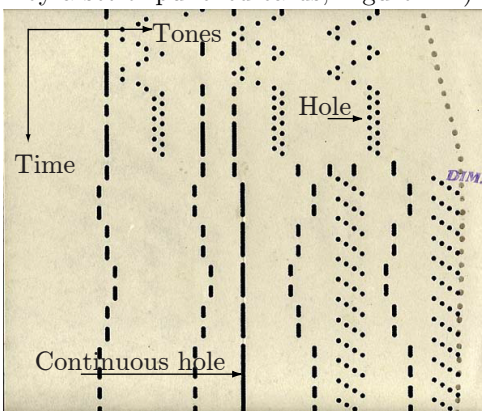
The following equalities add to the interest of 2: $2 + 2 = 2 \times 2 = 2^2$, $2 = 2!$, and $2 \uparrow\uparrow 2 = 2 \uparrow\uparrow\uparrow 2 = 2 \uparrow\uparrow \ldots \uparrow 2 = 4$ (where the $\uparrow$ notation is described on Page 66).

The integer 2 forms the basis of the sequence of square numbers. These are figurate numbers of the form $S_n = n^2$, for $n \geq 0$. Thus, 0, 1, 4, 9, 16, 25, 36, 49,... (see Figure 2.5 and sequence A000290 in [Sloane 08]).

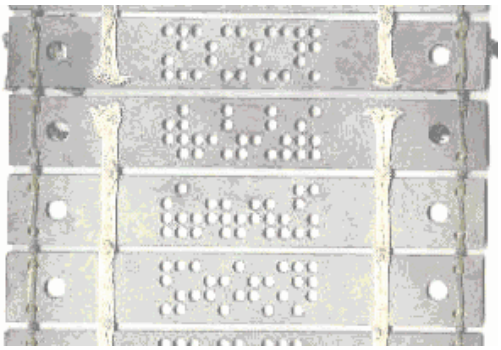
(Figurate numbers are those that can be represented by a regular geometrical arrangement of equally-spaced points. Figure 2.5 illustrates a few examples.)

The importance of 2 is illustrated by the use of words such as dual, duel, couple, double, pair, dichotomy, duplex, duet, dyad, duplicity, and twins in modern languages.

In religion, the ten commandments were given on two stone tablets. In biology, there are two sexes. Most important, in modern technology, mechanical looms, player pianos, and computers are based on binary (base-2) numbers. (A player piano is driven by a perforated paper tape. Each hole punched in the tape represents a tone to be played, whereas the absence of a hole signifies silence. Old mechanical looms were similarly driven by a set of punched cards, Figure 2.4.)



Player piano roll



Mechanical loom cards

Figure 2.4: Player Piano Paper Roll and Mechanical Loom Punched Cards.

The number 2 also represents the simplest symmetry. Drop a little ink on a sheet of paper, fold the paper once, press, and open. The resulting pattern exhibits a twofold

symmetry which forms the basis of the Rorschach inkblot test, a method of psychological evaluation.

3. There isn't much to say about this old acquaintance, the integer 3. Nevertheless, it isn't completely boring. It is the first odd prime, it is associated with an interesting divisibility test, and it is the number of points in a triangle, itself an important geometric figure.

Given an integer, such as 1632, it is easy to check whether it is divisible by 3. Simply add its digits and check divisibility of the sum. Thus, $1+6+3+2=12$ and $1+2=3$, indicating that 1632 is divisible by 3. Addition is commutative, so we conclude that any integer that consists of the digits 1, 6, 3, and 2, such as 2361, is also divisible by 3.

◇ **Exercise 2.4:** Prove this divisibility test.

◇ **Exercise 2.5:** Show that integers of the form $100\ldots061600\ldots01$ are not primes regardless of the number of zeros.

A triangle is defined by three independent points (i.e., three points not on the same line). Thus, a triangle is a polygon with the smallest number (3) of edges. It is always flat, because three distinct points are always in the same plane. Triangles have many interesting and useful properties, one of which will be mentioned here. Imagine a triangle whose three edges are connected with hinges. In spite of this flexibility, the triangle will resist attempts to change its shape; it is a durable geometric figure. (Three points not on the same line also define a unique circle.)

A curiosity: Take any 3-digit integer *abc*, append the same three digits to obtain the 6-digit integer *abcabc*, divide it by 7, then by 11, and finally by 13. You'll be surprised to see that the final result is the original 3-digit integer *abc*. (For those who are not afraid of algebra, this is because $7\times11\times13=1001$. Also see Page 57 for the divisibility test for 7.)

The sequence of triangular numbers is well known. It consists of the partial sums of the divergent series $1+2+3+4+5+\cdots$. Thus, the first few such numbers are 1, 3, 6, 10, 15, 21, 28, ... In general, The nth triangular number T_n is given by $n(n+1)/2$. These numbers have many interesting and beautiful properties, two of which are mentioned here:

The sums of consecutive triangular numbers form the following patterns

$$T_1+T_2+T_3=T_4,$$
$$T_5+T_6+T_7+T_8=T_9+T_{10},$$
$$T_{11}+T_{12}+T_{13}+T_{14}+T_{15}=T_{16}+T_{17}+T_{18},$$

and so on.

The triangular numbers increase as we move along their sequence, which is why their inverses decrease. It can be shown that the infinite sum of those inverses converges. Thus

$$1+1/3+1/6+1/10+1/15+1/21+1/28+\cdots=2.$$

In religion, there is the Christian Holy Trinity. In science, the proton and neutron each consist of three quarks. The world that we perceive is three-dimensional (in special

relativity, time is added as a 4th dimension, but is different from the three dimensions of space). In biology, each codeword of the genetic code consists of three nucleotides.

The triple redundancy concept, introduced in [Von Neumann 56], is based on having three redundant copies of any important device, circuit, or instrument that may fail. The following quotations refer to this concept.

> The Ramans do everything in threes.
> —Arthur C. Clarke, *Rendezvous With Rama*, 1973.
>
> What I tell you three times is true.
> —Lewis Carroll, *The Hunting of the Snark*, 1876.

4. The integer 4 is the smallest composite number. Its prime factors are 2 and 2, so 4 is also the smallest composite number that equals the sum of its prime factors. It is also the smallest square prime (because $4 = 2^2$ and 2 is a prime) and is the only even number that is a square prime.

A decimal integer is divisible by 4 if its two least-significant digits (considered as a two-digit integer) are divisible by 4. Thus 167,036 is divisible by 4 because 36 is. This is the divisibility test for 4, because all the digits to the left of the rightmost two are multiples of the form 100, 1000, and so on, and are all divisible by 4.

A quadrilateral is a plane figure with four edges. The most common quadrilaterals are the square and the rectangle. Given four two-dimensional points, no three of which are collinear, they uniquely define a hyperbola. Every positive integer is the sum of at most four integer squares (Table 2.7).

The number 4 plays an essential role in the four-color theorem. This important result states that four colors are enough to color regions on a map, such that no adjacent regions have the same color. (Two regions are considered adjacent if they share a segment, not just a point.)

In biology, DNA consists of the four nucleobases adenine, guanine, cytosine, and thymine. Physics tells us that there are four fundamental forces in nature (gravity, electromagnetism, the weak nuclear force, and the strong nuclear force). Many vehicles have four wheels and many pieces of furniture have four legs. In English, the phrase "four-letter word" indicates a swear word. Also, the word "four" consists of four letters and is the only English name of a digit with this property. (See also the fun sequence on Page 304.) The ancient Greeks believed that fire, air, water, and earth were the basic constituent elements of the world. There are four directions, four annual seasons, four corners of the world, and four suits of playing cards.

See [Penelope 09] for more thoughts on the number 4.

The digit 4 has an interesting property in the English language, a property that makes it an attractor. Choose any number, integer, irrational, real, or complex, and spell it in English. If its English name consists of n letters, then spell the English name of the number n. When this process is repeated, it always results in 4, because 4 is the only number n whose English name consists of 4 letters.

5. We are now halfway between 0 and 10. Five is a prime number, but its main claim to mathematical fame is being the only prime that ends with a 5. An integer is divisible by 5 if its least-significant digit is 0 or 5. Thus, integers that end with 5 are divisible by 5 and are therefore composite.

In geometry, there are five Platonic solids, the tetrahedron, cube (or hexahedron), octahedron, dodecahedron, and icosahedron.

A polygon with five sides is a pentagon. This shape gives rise to the little-known pentagonal numbers p_n. Figure 2.5 illustrates the geometric meaning of those numbers. It can be shown that the nth pentagonal number is given by $p_n = n(3n-1)/2$ for $n \geq 1$ (see sequence A000326 in [Sloane 08]).

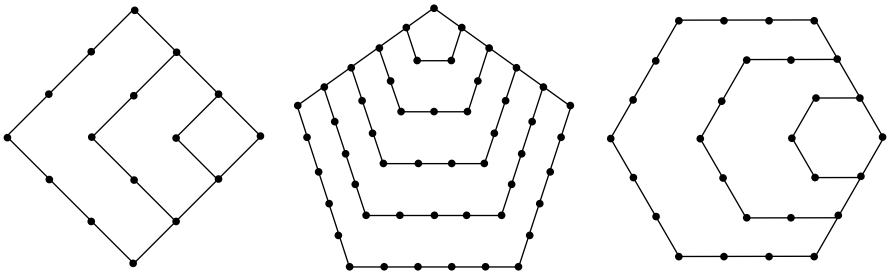


Figure 2.5: Several Figurate Numbers.

In biology, almost all amphibians, reptiles, and mammals with fingers or toes have five of them on each extremity. In a two-body astronomical system, such as a planet around a sun or a moon around a planet, there are five points where a small object is at equilibrium. Once placed at such a point, the object will remain in orbit. It will not escape and will not fall toward either of the bodies. These points are referred to as the Lagrangian points of the system. (Only two of the five Lagrangian points are stable, and they exist only if the masses of the two astronomical bodies satisfy certain conditions.) The standard musical notation is based on a musical staff made of five horizontal lines.

In technology, many devices—such as computer keyboards, test instruments, and telephones—have numeric keypads where the 5 key has a raised dot or bar to make dialing easier for the visually handicapped.

Digression. The following Mathematica code shows a beautiful relation between the integer 5 and the all-important, all-pervading constant π. This is one of the many unexpected results that are often discovered by lovers of mathematics and are enjoyed by many. It is illustrated in [Pi5 20].

```
a=5.;
Do[{Print[{a,Sin[(1./a) Degree]}],a=10a+5},{i,7}]
```

This code iterates several times, generates numbers of the form 5, 55, 555, ..., and computes the sine of their inverses. Surprisingly, the results approach π.

Like many other unexpected mathematical surprises and tricks, the explanation of this result is simple and is based on the following facts:

1. The number 1/5, the inverse of 5, contains the digits 18 along with several zeros. Thus, for example, $1/555555 = 0.0000018000\cdots \approx 1.8 \times 10^{-6}$.

2. For small angles θ, the trigonometric function sine satisfies $\sin\theta \approx \theta$.

3. The relation between degrees and radians is $360^\circ = 2\pi$ radians, which is why one degree equals $\pi/180$ radians.

With this in mind, the following expression explains the relation between 5 and π

$$\sin\left[\frac{1}{555555}\right] = \sin(1.8\times 10^{-6}) \text{ degree} = \sin\left[1.8\times 10^{-6}\times\frac{\pi}{180}\right] \text{ radian}$$
$$\approx 1.8\times 10^{-6}\times\frac{\pi}{180} = \pi\times 10^{-8}.$$

End of Digression.

6. Moving along, we now meet the integer 6. Six is a perfect number because it equals the sum of its divisors, $6 = 1 + 2 + 3$. It is also the smallest perfect number. Six is the only number that is both the sum and the product of three consecutive positive integers. The hexagonal numbers h_n (sequence A000384 in [Sloane 08]), are defined similar to pentagonal numbers (Figure 2.5) and it can be shown that $h_n = n(2n-1)$.

The divisibility test for 6 is simple. An integer is divisible by 6 if it divisible by 2 and by 3. Thus, for example, 738 is divisible by 2 (because it is even) and also divisible by 3 (because $7 + 3 + 8 = 18$).

In geometry, a cube has six faces, which is why 6 is the largest number on a standard die. Also, hexagons can tile the plane. In biology, the cells of a beehive honeycomb are hexagons as also are snowflakes (Figure 2.6). The standard model of particle physics includes six types of leptons and six types of quarks. In consumer products, six is the number of cans of soda or beer in a six-pack. In technology, six is the number of dots in a Braille cell. "Sex-" is a Latin prefix meaning "six" which is why we have words such as sextet and sextuplet.



Figure 2.6: The Number 6 in Biology.

◇ **Exercise 2.6:** What are other occurrences of six in nature, in addition to beehives?

7. This integer has less claim to fame than its predecessors. As an integer, its main interest is the little-known fact that it is the only integer less than 15 that cannot be represented as the sum of the squares of only three integers (it requires four squares). In 1770, Lagrange proved the four-square theorem which states that every positive integer is a sum of at most four integer squares. This is illustrated in Table 2.7 for the first 16 positive integers. The table lists each integer as the sum $a^2 + b^2 + c^2 + d^2$ and it is easy to see that 7 equals the sum $2^2 + 1^2 + 1^2 + 1^2$ and that the next integer that requires four squares is $15 = 3^2 + 2^2 + 1^2 + 1^2$.

The divisibility test for 7 is not trivial. Given an integer a, subtract twice its least-significant digit from the remaining digits (alternatively, add five times the least-significant digit to the sum of the remaining digits). If the result is divisible by 7, then

i	a	b	c	d	i	a	b	c	d
1	1				9	3			
2	1	1			10	3	1		
3	1	1	1		11	3	1	1	
4	2				12	2	2	2	
5	2	1			13	3	2		
6	2	1	1		14	3	2	1	
7	2	1	1	1	15	3	2	1	1
8	2	2			16	4			

Table 2.7: The Four-Square Theorem.

a is too. Thus, 357 yields $35-2\times 7=21$ or $35+5\times 7=70$, implying that it is divisible by 7.

The integer 7 is also the base of the figurate heptagonal numbers that are defined by $h_n=(5n^2-3n)/2$ for $n\geq 1$. The first few members of this sequence are 1, 7, 18, 34, 55, 81, 112, 148, 189, 235, 286, ... (sequence A000566 in [Sloane 08]).

Thus, seven is not very important or interesting in mathematics, but is common and important in other areas of science and everyday life. Here are just a few examples.

Given a six-sided die, the sum of the numbers on each pair of opposite faces of the die is 7.

In physics, as well as in poetry, there are seven colors of the rainbow, red, orange, yellow, green, blue, indigo, and violet. In technology, any decimal digit can be displayed by lighting some or all of seven segments. The seven-segment display has been common in computers and instruments since the 1970s. There are also seven logic gates, NOT, AND, OR, NAND, XOR, NOR, and XNOR. In history, mythology, and religion, the world was created in seven days, the ancient world had seven wonders, the week has seven days, there are seven deadly sins, the number seven was considered magical, Rome was built on seven hills, and a cat is said to have seven lives. There are many references to seven in the *Harry Potter* novels. In music, there are seven notes in the traditional Western Major Scale.

For more examples of the importance of seven see [seven magazine 09].

> As I was going to St. Ives, I met a man with seven wives. Each wife has seven kids, each kid has seven cats, each cat has seven kittens, each kitten had seven fleas. Fleas, kittens, cats, kids, wives, how many were going to St. Ives?
>
> —A Traditional English Language Nursery Rhyme.

8. In computers, the octal number system, which is based on 8, used to be popular with old computers (whose memories often featured words lengths that were multiples of 6) because the relation $8=2^3$ implies that each octal digit equals exactly three bits, and it is easy to convert between binary and octal. Thus, short octal numbers were sometimes used instead of the much longer binary numbers. Also, the smallest addressable unit in most modern computer memories is the byte, which consists of eight bits (but may sometimes have other lengths, a feature that introduces confusion).

An 8-sided polygon is an octagon. Figurate numbers based on octagons are called octagonal. The first few such numbers are 1, 8, 21, 40, 65, 96, 133, 176, 225, 280, 341,

and 408 (sequence A000567 in [Sloane 08]). The nth octagonal number has the form

$$n^2 + 4\sum_{k=1}^{n-1} k.$$

Here is the divisibility test for 8. Given an integer a, concentrate on its three least-significant digits, because everything to their left is always divisible by 8. Add the least-significant digit to twice the remaining digits. If the result is divisible by 8, then a is too. For example, given $ddd192$, we add $2 \times 19 + 2 = 40$. If we are not sure about the divisibility of 40, repeating the same process yields $2 \times 4 + 0 = 8$, which shows that 192 is divisible by 8. The proof of this divisibility test is simple and is similar to the one described in the solution to Exercise 2.7.

The term "figure eight" is common in sports (skating and air races immediately come to mind) and in knots (figure eight knots around cleats or pins are used to belay heavy loads).

In particle physics, the eightfold way [eightfold 00] is used to classify sub-atomic particles.

In biology, all spiders (in general, all arachnids), have eight legs. The octopus and its relatives have eight arms. In architecture, gazebos and certain pagodas tend to have eight sides. In music, the interval between two notes x and y with frequencies f and $2f$, is called an octave, because on a standard scale there are eight notes between x and y (including x and y). In technology, many street signs, such as the well-known stop sign, have eight sides.

> The number 8 plays a special role in mathematics due to octonions, an 8-dimensional number system where one can add, multiply, subtract and divide, but where the commutative and associative laws for multiplication—$ab = ba$ and $(ab)c = a(bc)$—fail to hold. The octonions were discovered by Hamilton's friend John Graves in 1843 after Hamilton told him about the quaternions. While much neglected, they stand at the crossroads of many interesting branches of mathematics and physics.
>
> —John Baez, 2008.

9. Nine is the largest decimal digit, and it is this simple property that makes it the darling of merchants, business people, and sellers of all kinds. In today's business world, round prices have all but disappeared. Nothing costs 10 pounds, Euros, or dollars, but 9.99 is a common price, as are 0.99, 24.99, 199.99, 549.99, and many others. The Preface tells us that many people are ignorant of mathematics or are afraid of it, and it is precisely those people who see a price tag of 15,999 and say "fifteen thousand, not bad. Last year I paid 16,000 for the same item." This property of nine makes it important and interesting in marketing, the economy, and everyday life in general.

Subsection 2.6 discusses the unusual real number $0.999\ldots$ (that is also denoted by $0.\bar{9}$, $0.\dot{9}$, and $0.(9)$) and proves in several ways that this number equals 1. This is true for other number bases, which is why the binary number $0.111\ldots_2$ equals 1. Similarly, any number that ends with an infinite number of 9's has a simpler, shorter representation. Thus, 17.5725 can also be written $17.5724999\ldots$.

The divisibility test of 9 is similar to that of 3. An integer is divisible by 9 if the sum of its digits is divisible by 9.

◇ **Exercise 2.7:** Show how this divisibility test can be proved in general (compare with Exercise 2.4).

People of my generation (and many younger ones) learned in school that there are nine planets in our solar system. In 2006, Pluto was demoted to a minor planet, leaving us with only eight planets.

A party trick. Take nine playing cards of any suit, labeled 1 through 9 (or nine domino pieces valued 1 through 9) and ask a friend to shuffle them. Spread them to form a nine-digit integer. Pretend to scrutiny this number thoroughly, and then claim that it is divisible by 9. Any calculator would prove you right. Challenge your friend to shuffle again, and repeat this trick several times. All those 9-digit integers would divide evenly by 9 because the sum of their digits is 45, which is divisible by 9.

The digit 9 as an attractor. The following property of 9 is a direct result of its being one less than the base 10 of the decimal numbering system. Start with a multi-digit number and subtract from it the sum of its digits. The difference will always be a multiple of 9. Repeating this process eventually results in 9, which is why this digit is often considered the attractor of the decimal system. A similar property exists for any number base. Here is an example of a three-digit decimal number abc. When subtracting the sum of digits from this value, the result is $100a + 10b + c - (a + b + c) = 99a + 9b$, which is obviously divisible by 9.

10. We have finally arrived at 10. The most-important attribute of 10 is that it is the basis of our number system (because we have ten fingers or digits), and as a result it is also the first 2-digit positive integer. Also, many monetary units (ruble, Euro, dollar, and mark) are divided into 100 parts (kopeks, cents, or pfennig), and there often are coins worth 10 parts.

This number has another property that makes it interesting (and perhaps also unique). The well-known factorial function is defined by $n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$ and it is easy to verify the identities $n! = n(n-1)!$ and $(n!)! = n!(n!-1)!$. Thus, integers of the form $(n!)!$ are products of the factorials of two consecutive integers. The only other known integer that can be written as such a product is 10, because $10! = 6!7!$.

An integer is divisible by 10 if its least-significant digit is 0.

The Biblical ten commandments are fundamental laws in both Judaism and Christianity. A set of ten items (such as ten years) is called a decade. A perfect score is sometimes referred to as a "10."



Figure 2.8: Image by MS Copilot, Sept. 2025.

11. There is not much to say about 11. Its divisibility test is slightly confusing. Scan the digits from right to left and perform an alternating sum, i.e., $a_0 - a_1 + a_2 - a_3 + \ldots$. If the result is divisible by 11, then the entire number is.

We have looked into zero through 11 and have found interesting features, attributes, and applications for each. What about other integers? Before listing interesting features of larger integers, let's look at a simple, funny proof of the following claim: Every integer is interesting!

The proof (see also Page 2) is simple. Assume that there is an uninteresting integer (or even several such integers). Denote the smallest of them by A. Thus, A is the lowest uninteresting integer, an attribute that by definition makes A interesting (perhaps even very interesting). The resulting contradiction proves that our assumption is wrong.

This "proof" can easily be extended to rational numbers. A rational number R can be written as the quotient a/b of two integers where b is nonzero. Since a and b are integers, they are interesting, so R can also be considered interesting.

This tongue-in-cheek proof is a paradox that stems from the vagueness of the words "interesting" and "dull." This topic is further discussed in [WikiInteresting 09].

The integers may be the "simplest" numbers, but even they have many surprising, interesting, and perhaps also fascinating properties. Here are two examples.

The sum of any n consecutive natural numbers equals the sum of the $n-1$ numbers that follow. Thus, $1+2=3$, $4+5+6=7+8$ and $9+10+11+12=13+14+15$.

The sum of any n consecutive odd integers divided by the sum of the n following odd integers is always 3.

We continue with several more interesting numbers.

1.618033.... This is the celebrated golden ratio (φ), the topic of Subsection 2.11.4. Here, we mention just a few interesting facts about this well-known number. The golden ratio is the solution of the quadratic equation $\varphi = 1 + 1/\varphi$. This surprising equation implies many attributes of φ as follows: (1) The inverse $1/\varphi$ is the difference $\varphi - 1$. (2) Writing this equation in the form $\varphi^2 = \varphi + 1$ shows us that the square of the golden ratio is the sum $\varphi + 1$. (3) The relation $\varphi = \varphi^2 - 1 = (\varphi - 1)(\varphi + 1) = (\varphi - 1)\varphi^2$ implies that $\varphi = 1/(\varphi - 1)$.

In addition to all its other beautiful and interesting properties, the golden ratio φ is the only real number that satisfies the three interesting relations above. See Subsection 2.11.4 for more information.

2.6854.... Khinchin's constant. Continued fractions are the topic of Section 2.5. Consider the simple continued fraction

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \cfrac{1}{\ddots}}}}}.$$

It defines a sequence (infinite for irrational numbers) of coefficients a_i which are specific to x. It comes as a surprise to realize that the limit of the geometric mean of these coefficients (almost) always converges to 2.6854520010.... Thus,

$$\lim_{n\to\infty} (a_1 a_2 a_3 \ldots a_n)^{1/n} = K.$$

This limit is called Khinchin's constant and is denoted by K_0. Discovered by Aleksandr Khinchin, this constant is defined as the limit of the product

$$K \stackrel{\text{def}}{=} \Pi_{r=0}^{\infty}\left[1+\frac{1}{r(r+2)}\right]^{\log_2 r}.$$

The first few digits of K are listed as `OEIS` sequence `A002210`, reference [Sloane 08].

Even though K_0 is known to many thousands of digits, it is not known if it is irrational, let alone transcendental. On the other hand, this constant applies to itself. The Mathematica statement

```
N[GeometricMean[ContinuedFraction[Khinchin, 5000]]]
```

produces 2.66235 (many more coefficients are needed for a more accurate value).

We can claim that Khinchin's constant K_0 "knows" about almost all numbers, which makes it one of the most important mathematical constants, and yet it is not widely known.

2.7 1828 1828 45 90 45.... This is the important Euler constant e. The notation here illustrates how easy it is to memorize its first 16 digits. This constant is normally known as the base of natural logarithms, but it is also important in financial calculations because it is the limit of the expression $(1+1/n)^n$ when n tends to infinity, which is used to compute compound interest. In addition to its interest and importance, e is an important part of the beautiful equation $e^{\pi i}-1=0$ and is also related to the trigonometric functions sin and cos by the elegant expression $e^{i\theta}=\cos\theta+i\sin\theta$. In calculus, the derivative of the function e^x is itself. Thus, the rate of change of e^x when x varies is e^x. A remarkable property!

A mnemonic for e. How to remember our important friend e to more decimal places than we will ever need? Just memorize 2 point 7 followed by birth year of Leo Tolstoy (1828), followed by birth year of Jules Verne (1828), followed by angles of isosceles right triangle (45, 90, 45), to obtain $e = 2.7|1828|1828|45|90|45\ldots$.

3.1415926.... This is π, the most famous, remarkable, and mysterious number. Start with an arbitrary circle, measure its circumference (c) and its diameter (d) and divide these numbers. The ratio c/d equals π regardless of the circle selected. Most grownups are used to this fact and take it for granted, but when a child hears of it for the first time, it comes as a surprise. Why should such property hold for all circles? There is no satisfying explanation. The ancients already knew, or suspected, that this ratio is constant and mathematicians throughout history tried to compute its value to more and more digits.

A remarkable property of π is that it appears in many infinite series. The following example is especially interesting

$$\frac{\pi^2}{6}=1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots.$$

◇ **Exercise 2.8:** Why is this example especially interesting?

Out of the vast number of applications of π, its use in popular culture, and its many appearances in science, engineering, and mathematics, I would like to mention the

beautiful and surprising video at [PiCollision 19]. It shows an unexpected connection between the digits of π and the number of perfect, elastic collisions between two ideal, frictionless blocks of matter.

12. There are curious properties that make the integer 12 interesting. One is the product $12 = 3 \times 4$, a pattern that continues with $56 = 7 \times 8$, and the other is that $12^2 = 144$ and $21^2 = 441$, curious indeed!

It has been pointed out that 12 would be better than 10 as the base of a number system because it is divisible by 2, 3, 4, and 6, in contrast with 10, which is only divisible by 2 and 5. Thus, in the decimal number system, the inverses of 2 and 5 are the short fractions 0.5 and 0.2, but all other inverses of decimal digits are longer fractions, such as 0.25, 0.333..., and $1/6 = 0.1666...$. A number system based on 12 (duodecimal) has 12 digits, and the inverses of 2, 3, 4, and 6 (i.e., the halves, thirds, quarters, and sixths) are short.

Most of the world uses the metric (or decimal) system of weights and measures, but a few countries still struggle with 12 inches to the foot and 16 ounces to the pound. Until 1971, England had a monetary system of 12 pence (d) to the shilling (s) and 20 shillings to the pound (£), in addition to units such as a crown (5s), half a crown (2s 6d), a florin (2s), half a sovereign (10s) and a guinea (21s). A statement such as "I only paid two pounds, six and nine" meant two pounds, six shillings, and nine pence.

> "I was born in 1943. The money used in our village was: farthing, Ha'penny, penny, thrupenny bit, sixpence, shilling, two bob bit, half crown, ten bob note, pound note and five pound note. The crown coin was limited. I don't think there was a five pound coin. I believe the guinea was, still is, just a value and not a coin or note."
>
> —John Curd.

16. The equation $a^b = b^a$ is simple and symmetric. Like any equation in two unknowns, it has infinitely many solutions, but this particular equation has only one integer solution, namely $a = 4$ and $b = 2$ (or $a = 2$ and $b = 4$). Thus, $4^2 = 2^4 = 16$.

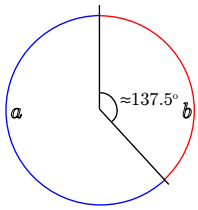
Computers use binary numbers and many current computer memories and devices are organized around units of eight bits, bytes. This is why the hexadecimal (base-16) number system is important. Binary numbers are simple because they employ just two digits, but the tradeoff is that these numbers tend to be long. The hexadecimal system is more complex, having the 16 symbols 0 through 9 and `A`, `B`, `C`, `D`, `E`, and `F`, but hexadecimal numbers are much shorter. In addition, it is easy to convert between binary and hexadecimal numbers, because a group of four bits can have 16 values. Thus, a byte can be divided into two groups of four bits each and thereby be expressed by two hexadecimal digits. The digits `2A` are equivalent to 0010|1001 and `F0` denotes the byte 1111|0000. Decimal numbers are also shorter than binary, but a decimal digit is equivalent to about 3.32 bits, which complicates conversion between decimal and binary.

37. Reference [37everywhere 24] shows that when we are asked to produce a random integer between 1 and 100, the most common choices are 3, 37, and 73. See Section 7.10 for an unexpected appearance of 37 in real life.

69. It is easy to verify that $69^2 = 4{,}761$, that $69^3 = 328{,}509$, and that between them these two numbers employ each of the ten decimal digits once. It seems that 69 is the only integer that has this interesting property.

◇ **Exercise 2.9:** What is the next element of the sequence one, seven, eleven, twenty one, seventy seven?

137.508. This irrational number is the little known golden angle.



The golden angle, the close cousin of the golden ratio, is constructed by partitioning a circle into two arcs, large a, and small b, such that $a/b = \phi$. This happens when the angle between the radii defining the arcs is approximately 137.5°, as illustrated in the figure. The golden angle is often found in plants where it is the angle between successive leaves. Natural selection favors plants that receive more sunlight, which is why the arrangement of leaves around a plant is important. The worse phyllotaxis occurs when leaves grow under or over one another, thereby blocking the sun.

The value of the golden angle is easy to compute (see Page 291). The circumference of a circle is always 2π and in our case $a+b$. The ratio of the arc subtended by b to the entire circumference is therefore

$$\frac{b}{2\pi} = \frac{b}{a+b} = \frac{a+b}{b}^{-1} = 1 + \frac{b}{a}^{-1} = \frac{1}{1+\phi} = \frac{1}{\phi^2}.$$

Thus, exactly ϕ^2 golden angles fit in the 2π circumference of the circle, which results in

$$\text{Golden angle} = \frac{2\pi}{\phi^2} \approx 2.39996 \text{ radian}, \text{or} \frac{360}{\phi^2} \approx 137.508.$$

108. This number equals $1^1 + 2^2 + 3^3$ and is also the interior angle, in degrees, of a regular pentagon.

153 is both a triangular and a hexagonal number. It also satisfies

$$153 = 1^3 + 5^3 + 3^3 = 1! + 2! + 3! + 4! + 5!$$

276. This is currently (in 2024) the smallest number whose aliquot sequence is not fully known due to the difficulty of factoring large numbers. See Page 284.

495. This integer is the basis of the Kaprekar process. This is an algorithm discovered (invented? originated?) in 1949 by Dattatreya Ramachandra Kaprekar, which delights many mathematicians (mostly amateurs). Take any three-digit integer whose digits are not all identical. Arrange the digits in ascending and descending order and subtract the smaller resulting integer from the larger. Notice that the result may have some identical digits. If the result has fewer than three digits, prepend zeros as necessary to end up with three digits. Repeat this process and you will always end up with 495, an integer that is transformed to itself by this process because $954 - 459 = 495$.

Repeating the same process with a four-digit integer results in 6174.

The integer 1,729 is known as the Hardy-Ramanujan number after a well-known anecdote told by the mathematician G. H. Hardy. In his essay *A Mathematician's Apology*, Hardy tells of a hospital visit to the Indian mathematician Srinivasa Ramanujan

> "I remember once going to see him when he was ill at Putney. I had ridden in taxi cab number 1729 and remarked that the number seemed to me rather a dull one, and that I hoped it was not an unfavorable omen. 'No,' he replied, 'it is a very interesting number; it is the smallest number expressible as the sum of two cubes in two different ways.' "



Figure 2.9: Image by MS Copilot, Sept. 2025.

When hearing this anecdote, the mathematician J. E. Littlewood responded by saying "every positive integer was one of Ramanujan's personal friends."

See [MathPages 09] for further discussion of 1,729 as an interesting or dull number.

◇ **Exercise 2.10:** Find out (by searching or by direct computations) [1] the two pairs the sum of whose cubes equals 1,729. [2] The two pairs of integers the sum of their 4th powers equals 635,318,657.

1093. This is a Wieferich prime. These are prime numbers p such that p^2 divides $2^{p-1} - 1$. So far, only two members of this sequence, 1093 and 3511, are known.

1661. This is Belphegor number of order 0. It is a prime number (the 1928th prime, to be precise), and it has the following properties:

1. The sum of its digits (20) is also the sum of the digits of its position (1928) in the sequence of primes. Other primes that exhibit this property are 131 (the 32nd prime) and 263 (the 56th prime).

2. It is (*i*) a palindromic prime. (*ii*) A naughty prime, a prime number that has more zero digits than nonzero digits. (*iii*) A beastly prime, a prime number that contains the string 666.

3. It is the first number in the Belphegor family, whose general member is

$$B(n) = 1\underbrace{00\ldots0}_{n}666\underbrace{00\ldots0}_{n}1.$$

3435 also equals $3^3 + 4^4 + 3^3 + 5^5$.

142,857. This apparently run-of-the-mill integer has two beautiful properties. Firstly, it is the periodic part in $1/7 = 0.142857\,142857\,142\ldots$. Secondly, its first six products are cyclic permutations of itself, Thus,

$$\begin{aligned} &142{,}857 \times 1 = 142{,}857, \quad 142{,}857 \times 2 = 285{,}714, \quad 142{,}857 \times 3 = 428{,}571, \\ &142{,}857 \times 4 = 571{,}428, \quad 142{,}857 \times 5 = 714{,}285, \quad 142{,}857 \times 6 = 857{,}142. \end{aligned}$$

See Page 164 for more about this and other cyclic numbers.

998,001. This number is of interest because its reciprocal `N[1/998001, 3100]` is a string of three-digit numbers. They are arranged in a strange, but nonrandom order, making it tricky to verify whether any is missing. The list of Figure 2.10 consists of 3100 digits, more than 1000 numbers.

73,939,133. This is a prime number with a very unusual property. If we remove its least-significant digits one by one, the remaining integers—7,393,913, 739,391, 73,939, 7,393, 739, 73, and 7—are all primes! Other primes with this property are known.

381,654,729. A curious integer. It consists of the digits 1 through 9 and its leftmost k digits are divisible by k (for example, 38,165 is divisible by 5). Reference [Lanier 09] shows how Susie Lanier cleverly employs logic and elementary algebra to determine this number.

$\mathbf{10^{3003}}$. This number has the familiar-sounding but little-known name millillion. See [Mufano 09] for ways to construct and name large and very large numbers. One such method, the up-arrow notation devised by Donald Knuth [Knuth 76], is described here.

The up-arrow notation. Modern mathematical notation often employs iteration. Thus, multiplication is iterated addition $a \times b = \underbrace{a + a + \cdots + a}_{b}$ and exponentiation is iterated multiplication $a^b = \underbrace{a \times a \times \cdots \times a}_{b}$. The up-arrow notation also uses iteration. The basic operation $a \uparrow b$ is defined as a^b. Its immediate extension $a \uparrow\uparrow b$ is defined as

$$a \uparrow\uparrow b = \underbrace{a \uparrow (a \uparrow (\ldots \uparrow a))}_{b} = a^{a^{\cdot^{\cdot^{\cdot^{a}}}}\Big\}b}.$$

Thus, $4 \uparrow 3 = 4^3 = 64$ and $4 \uparrow\uparrow 3 = 4 \uparrow (4 \uparrow 4) = 4^{4^4} = 4^{256}$. Notice that 4^{4^4} is to be computed from right to left, thus $4^{(4^4)}$.

Other examples of double up-arrows are

$$3 \uparrow\uparrow 2 = 3^3 = 27, \quad 3 \uparrow\uparrow 3 = 3^{3^3} = 3^{27}, \quad 3 \uparrow\uparrow 4 = 3^{3^{3^3}} = 3^{3^{27}}.$$

The definition of triple up-arrow is now obvious $a \uparrow\uparrow\uparrow b = \underbrace{a \uparrow\uparrow (a \uparrow\uparrow (\ldots \uparrow\uparrow a))}_{b}$.

```
1.00200300400500600700800901001101201301401501601701801902002102202302\
40250260270280290300310320330340350360370380390400410420430440450460470\
04804905005105205305405505605705805906006106206306406506606706806907000\
71072073074075076077078079080081082083084085086087088089090091092093090\
40950960970980991001011021031041051061071081091101111121131141151161170\
11811912012112212312412512612712812913013113213313413513613713813914010\
41142143144145146147148149150151152153154155156157158159160161162163160\
41651661671681691701711721731741751761771781791801811821831841851861870\
18818919019119219319419519619719819920020120220320420520620720820921020\
11212213214215216217218219220221222223224225226227228229230231232233230\
42352362372382392402412422432442452462472482492502512522532542552562570\
25825926026126226326426526626726826927027127227327427527627727827928020\
81282283284285286287288289290291292293294295296297298299300301302303300\
43053063073083093103113123133143153163173183193203213223233243253263270\
32832933033133233333433533633733833934034134234334434534634734834935030\
51352353354355356357358359360361362363364365366367368369370371372373370\
43753763773783793803813823833843853863873883893903913923933943953963970\
39839940040140240340440540640740840941041141241341441541641741841942040\
21422423424425426427428429430431432433434435436437438439440441442443440\
44454464474484494505145245345445545645745845946046146246346446546646700\
46846947047147247347447547647747847948048148248348448548648748848949040\
91492493494495496497498499500501502503504505506507508509510511512513510\
45155165175185195205215225235245255265275285295305315325335345355365370\
53853954054154254354454554654754854955055155255355455555655755855956050\
61562563564565566567568569570571572573574575576577578579580581582583580\
45855865875885895905915925935945955965975985996006016026036046056066070\
60860961061161261361461561661761861962062162262362462562662762862963060\
31632633634635636637638639640641642643644645646647648649650651652653650\
46556566576586596606616626636646656666676686696706716726736746756766770\
67867968068168268368468568668768868969069169269369469569669769869970070\
01702703704705706707708709710711712713714715716717718719720721722723720\
47257267277287297307317327337347357367377387397407417427437447457467470\
74874975075175275375475575675775875975607617627637647657667677687697707\
71772773774775776777778779780781782783784785786787788789790791792793790\
47957967977987998008018028038048058068078088098108118128138148158168170\
81881982082182282382482582682782882983083183283383483583683783883984080\
41842843844845846847848849850851852853854855856857858859860861862863860\
48658668678688698708718728738748758768778788798808818828838848858868870\
88888989089189289389489589689789889990090190290390490590690790890991090\
11912913914915916917918919920921922923924925926927928929930931932933930\
49359369379389399409419429439449459469479489499509519529539549559569570\
95895996096196296396496596696796896997097197297397497597697797897998090\
81982983984985986987988989990991992993994995996997998999000001002003004\
50060070080090100110120130140150160170180190200210220230240250260270280\
02903003103203303403 *10^-6
```

Figure 2.10: 1000 Three-Digit Integers.

and

$$a \uparrow\uparrow\uparrow\uparrow b \stackrel{\text{def}}{=} a \uparrow^4 b = \underbrace{a \uparrow^3 (a \uparrow^3 (\ldots \uparrow^3 a))}_{b}.$$

Thus, for example,

$$3 \uparrow^4 3 = 3 \uparrow^3 (3 \uparrow^3 3) = 3 \uparrow^3 (3 \uparrow^2 (3 \uparrow^2 3)) = 3 \uparrow^3 (3 \uparrow^2 27) = \ldots.$$

It is easy to see how this notation is extended to any number of consecutive up-arrows and how it provides very compact notation for extremely large numbers.

The integer 2 has the unique property that $2 + 2 = 2 \times 2 = 2^2 = 2 \uparrow\uparrow 2 = 2 \uparrow^n 2$ for any value of n.

$\mathbf{95^{10^6}}$. This unimaginably vast number gives us an idea of the size of the ultimate library. A book may have artistic, scientific, or literary value, but reduced to its essentials, it is a string of characters. A typical 300-page book may consist of about one million characters, letters, digits, and punctuation marks. A book may also contain mathematical expressions and figures, and the number of distinct characters depends on the language, but let's simplify our task and consider books that consist of a million characters drawn from a set of 95 distinct characters (this is the number of printable ASCII codes, including a blank space). Under these restrictions, total number of possible books is $\mathbf{95^{10^6} \approx 4 \times 10^{1,977,723}}$, a vast (close to two-million-digits-long) but as far from infinity as any other number.

Tablets were replaced by scrolls, scrolls were replaced by books, and now we scroll through books written on tablets.

A little thinking shows that the vast majority of these books would be complete or partial gibberish. Still, the number of meaningful books would be much greater than the number of elementary particles in the observable universe (a number which is estimated at about 10^{78} to 10^{82}). An interesting conclusion is that, given enough time, *all* these meaningful books might be written throughout history, during a long but finite interval of time.

A large integer. Here is an example of another large number that is easy to compute in principle, but impossible to compute with current technology. Start with the integers from 1 to 10^{100}. In this interval select the ones that are primes and compute their median. (If a sequence consists of an odd number of items, its median is the middle item. If it consists of an even number of items, its median is the average of the two items at the center of the sequence. The prime numbers are odd, so the average of two primes is an integer.)

Our sequence of primes starts with 2, 3, 5, and 7 and ends with

$$\underbrace{999\ldots99}_{97}203.$$

It consists of very large integers, but such integers can be computed by current software. The problem is the length of the sequence. Applying the prime number theorem to this case yields

$$\pi 10^{100} \approx \frac{10^{100}}{\ln(10^{100})} \approx \frac{10^{98}}{2.3} \approx 10^{97}.$$

This number is simply too large for any current computer to perform this computation (to compute the elements of the sequence in order to find the middle ones) in any meaningful period of time.

◇ **Exercise 2.11:** We have looked at many integers that posses interesting features. Many mathematics buffs claim that every integer belongs in this category, and no integer is boring. Search the mathematics world for an interesting property of the integer 196.

Self-describing numbers. A self-describing number is one that describes its own content in some way. For example, the integer 10,213,223 is self-describing because it consists (when viewed as 10|21|32|23) of one zero, two 1's, three 2's, and two 3's. The smallest self-describing number is 22. Self-describing numbers are easy to generate. Start with any positive integer, describe it, then describe the resulting integer, and repeat until the result is self-describing. Thus, starting with 101, we go through the sequence 1021, 102112, 103122, 10212213, 10313213, and end up with 10311233.

◇ **Exercise 2.12:** Starting with 10, evolve it in 10 steps to obtain a self-describing number.

The following "look and say" (or show and tell) sequence is an elegant variation on this concept, by John Conway (see also Page 303).

one
one one
two ones
one two and one one
one one, one two, two ones
three ones, two twos, one one
one three, one one, two twos, two ones

Tables 2.11 and 2.12 list beautiful pyramid patterns obtained by simple arithmetic operations on integers. Figure 2.13 illustrates unexpected relations between odd numbers and squares.

Another oddity of the odd numbers is the observation that the first odd number, 1, equals 1^3, the sum of the next two odd numbers $3+5$ equals 2^3, the sum of the next three odd numbers $7+9+11$ equals 3^3, and so on.

Here is a surprising fact about multi-digit integers. Take an n-digit integer whose digits are not all equal. Randomly select any of the $n!$ permutations of the digits, and subtract the permutation from the original number. If the difference is also a multi-digit number, add its digits. Repeat adding digits until a single-digit integer is obtained. This digit will always be 9. As an example, start with 108,473, subtract the permutation 470,813, to obtain (in absolute value) 362,340, add its digits to get 18, and add $1+8=9$. The world of mathematics is not just infinite; it is also full of unexpected results.

These relations and many others can be found at [numeropedia 09].

◇ **Exercise 2.13:** (Tongue in cheek). At school, students learn the meaning of percents and the fact that 100% of something equals that thing. In the school of life, however, a person learns (often from bitter experience) that there can be more than 100%. Interest rates, for example, are not limited to 100% and our leaders often ask us to produce a 101% effort. To gain an insight into the meaning of 101%, try the following fun experiment. Assign the integers 1 through 26 to the 26 letters A through Z. Given a word, replace its letters with their numeric values and add. It is now easy to verify that the sum of values for the word HARDWORK is $8+1+18+4+23+15+18+11=98$, while the code of KNOWLEDGE is $11+14+15+23+12+5+4+7+5=96$. If we consider these sums percentages, then hardwork is equivalent to giving 98% of yourself, knowledge is equivalent to 96%, and ATTITUDE equals $1+20+20+9+20+21+4+5=100\%$. Your task is to find the magic word or phrase that produces 101%.

$$
\begin{array}{rr}
1 \times 8 + 1 = 9 & \\
12 \times 8 + 2 = 98 & 9 \times 9 + 7 = 88 \\
123 \times 8 + 3 = 987 & 98 \times 9 + 6 = 888 \\
1234 \times 8 + 4 = 9876 & 987 \times 9 + 5 = 8888 \\
12345 \times 8 + 5 = 98765 & 9876 \times 9 + 4 = 88888 \\
123456 \times 8 + 6 = 987654 & 98765 \times 9 + 3 = 888888 \\
1234567 \times 8 + 7 = 9876543 & 987654 \times 9 + 2 = 8888888 \\
12345678 \times 8 + 8 = 98765432 & 9876543 \times 9 + 1 = 88888888 \\
123456789 \times 8 + 9 = 987654321 & 98765432 \times 9 + 0 = 888888888
\end{array}
$$

Table 2.11: Numerical Pyramids.

$$
\begin{array}{rr}
1\ 9 + 2 = 11 & 1 \times 1 = 1 \\
12\ 9 + 3 = 111 & 11 \times 11 = 121 \\
123\ 9 + 4 = 1111 & 111 \times 111 = 12321 \\
1234\ 9 + 5 = 11111 & 1111 \times 1111 = 1234321 \\
12345\ 9 + 6 = 111111 & 11111 \times 11111 = 123454321 \\
123456\ 9 + 7 = 1111111 & 111111 \times 111111 = 12345654321 \\
1234567\ 9 + 8 = 11111111 & 1111111 \times 1111111 = 1234567654321 \\
12345678\ 9 + 9 = 111111111 & 11111111 \times 11111111 = 123456787654321 \\
123456789\ 9 + 10 = 1111111111 & 111111111 \times 111111111 = 12345678987654321
\end{array}
$$

$$
\begin{array}{c}
296 \times 333667 = 98765432 \\
2996 \times 33336667 = 99876654332 \\
29996 \times 3333366667 = 99987666543332 \\
299996 \times 333333666667 = 99998766665433332 \\
2999996 \times 33333336666667 = 99999876666654333332 \\
29999996 \times 3333333366666667 = 999999876666666543333332 \\
299999996 \times 333333333666666667 = 9999999876666666543333333332 \\
2999999996 \times 33333333336666666667 = 99999999876666666654333333333332 \\
29999999996 \times 333333333336666666667 = 999999999876666666666543333333333332 \\
299999999996 \times 3333333333336666666667 = 9999999999876666666666543333333333332
\end{array}
$$

Table 2.12: More Numerical Pyramids.

$$
\begin{array}{c}
1 + 3 = 4 = 2^2, \\
1 + 3 + 5 = 9 = 3^2, \\
1 + 3 + 5 + 7 = 16 = 4^2, \\
1 + 3 + 5 + 7 + 9 = 25 = 5^2.
\end{array}
$$

Figure 2.13: Odd and Square Numbers.

> God made the integers; all else is the work of man.
> —Leopold Kronecker.

The above list of interesting integers becomes the tip of an infinitely big iceberg when we consider the various types of numbers that have been proposed and studied over the centuries. Here are three short lists of both common and rare types of numbers:

1. The elementary types, such as prime, integer, even, rational, irrational, real, constructible (those that can be made with a ruler and compass), algebraic (solutions to algebraic equations), transcendental (not a solution, but can be computed with an algorithm), non-computable (those that can be defined only by writing their digits explicitly), normal, surreal, complex, and quaternions.

An example of a non-computable number is Chaitin's constant Ω (the probability that a random program will halt) discovered and defined by Gregory Chaitin [WikiOmega 24].

An example of a normal number, one whose infinite sequence of digits is distributed uniformly, is Champernowne constant, the concatenation of all the natural numbers. Thus

$$C_{10} = 0.12345678910111213141516\ldots_{10} = 0.11011100101110111\ldots_{2}\,.$$

This number was proposed by David Champernowne in 1933.

◇ **Exercise 2.14:** Search the mathematical literature for another known normal number.

2. Well-known number types such as polygonal, Fibonacci, factorials, factorions, perfect, twin, amicable, oblong, sociable, Granville, twin primes, trigonometric, and computable.

A factorion is an integer which equals (in a certain number base) the sum of the factorials of its digits. Thus, 145 is a factorion because $1!+4!+5! = 145$. Other examples of decimal factorions are 1, 2, and 40,585. Notice that this property of integers depends on the number base. An integer which is a factorion in a base b may not be such in other number bases.

3. Rarely used numbers that are often defined by various relations. The following examples are mostly drawn from youtube channel `numberphile`: practical, witness, quasiperfect (or weird), dungeon, kissing, hyperreal, boring, cyclic, illegal, spaghetti, vampire, brown, deficient, abundant, hyperperfect, superperfect, definable, almost integer, and hypercomplex.

Normal numbers.

A simply normal number is a real number with an infinite sequence of digits where each digit appears with the same probability. A normal number is a simply normal number where every string of n digits appears with a probability of 10^{-n}. In base 10, for example, each of the ten digits of a normal number shows up about 10% of the time, each of the 100 pairs of digits shows up about 1%, and so on. A consequence of this definition is that every digit in the current number base must appear in the number. Notice that normality may not be absolute and may depend on the number base used. A number may be normal in one number base and not normal in other bases.

Every finite strings of digits appears in the representation of a normal number, which is why every literary work (including this book, as well as books made entirely of wrong facts), every image, and every video that have been made or will be made are included there. Given the time to search and find an as-yet-unknown video, can make you rich and famous.

Mathematically-savvy people often ask whether π is normal and whether every sequence of digits can be found in its decimal representation. The answers to these questions are still unknown. Currently, in 2025, more than 22 trillion digits of π are known, and it is easy to verify that all short strings of decimal digits (it is unclear up to what length) are indeed included among the known digits, but nothing is known about long strings.

A common example of a simply normal number is the binary $0.010101\ldots$. It contains the two binary digits and each appears exactly 50% of the time. Also, each of the strings 01 and 10 appears with frequency 0.5, but the bitstrings 00 and 11 never appear. As a decimal number, however, this number contains only two of the ten digits and is therefore not normal.

A number that is normal in every number base (base 2 and higher) is absolutely normal, and a number that is not normal in every base is absolutely abnormal. It is hard to find examples of such numbers, except that the rational numbers are not normal. The representation of such a number, in any base, eventually repeats, which is why many strings of digits appear only rarely or are completely missing.

It is known that almost all real numbers (except a set of measure zero) are normal, but the set of reals that are not normal is uncountably infinite.

Single Digit Representations of Natural Numbers

1. You are given a positive integer a between 0 and 1000, and a single nonzero digit n. Your task is to represent a as a mathematical expression using just n, the four arithmetic operations, parentheses, and exponentiation. Here are some examples:

$$544 = 2(2^{2\cdot(2+2)}2^{2+2}), \quad 441 = (5+5+55/5)^{(5+5)/5}.$$

2. Similarly, given an integer between 0 and 11111, represent it twice as a pandigital number, where each decimal digit appears exactly once. In one representation the digits are in ascending order and in the other they are in descending order. In addition to the four arithmetic operations, parentheses, and exponentiation, digits can be concatenated. Examples are:

$$6109 = 12^3 + 4 + 56\cdot 78 + 9 = 98(7\cdot 6 + 5\cdot 4) + 32 + 1,$$
$$3799 = 12 + 3 + 4 + 5(6+78)\cdot 9 = (9 + 8\cdot 7)\cdot 6\cdot 5 + 43^2\cdot 1.$$

These feats of computations were performed manually by Inder Jeet Taneja and made available by him as long lists in references [arxiv1502.03501] and [arxiv1302.1479]. There was, however, one missing entry in these lists, the pan digital ascending representation of 10,958. As can be expected of mathematicians, this gap did not remain empty for long. In 2017, the well-known recreational mathematician Matt Parker showed that the expression $1\cdot 23((4\cdot 5\cdot 6)||7+8)\cdot 9$ (where $||$ stands for concatenation) computes

to 10,958, thereby completing the heroic task started by Dr. Taneja. See reference [numberphile 17].

Chances are that many new types of numbers will be identified in the future, and this because there are infinitely many numbers of the elementary type, and because so many numbers have infinitely many digits. It is no wonder that such infinities of choice produce so many different types of numbers and relations between numbers.

This chapter started with a discussion of integers and their beautiful, interesting, and unexpected properties and relations. We continue our exploration of beauty, surprise, and fun in numbers by looking at other types of numbers.

2.3.1 Primes, composite, and their cousins

A prime number is an integer that has only itself and 1 as divisors. Prime numbers are important in mathematics because they are the building blocks from which other integers are constructed. This is summarized by the prime number theorem (also known as the fundamental theorem of arithmetic) which states that an integer can be factored into a product of prime numbers in one way only. For example, the prime factorization of 5040 is $2^4 \times 3^2 \times 5 \times 7$.

We use the number 10 as the base of our number system because we have 10 fingers, but mathematicians know that 12 is a better base of a number system, because it is divisible by 1, 2, 3, 4, 6, and itself, whereas 10 has only the four divisors 1, 2, 5, and itself. This raises the question of whether there are numbers better than 12 to base number systems on; whether highly-composite numbers exist, what makes them highly composite, and how to define them rigorously.

History teaches us that the Greek philosopher Plato had already realized that the number 5040 has many divisors. In his book *Laws* he used this to propose a way to organize a city according to some of those divisors. Plato perhaps even knew that those 60 divisors are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 18, 20, 21, 24, 28, 30, 35, 36, 40, 42, 45, 48, 56, 60, 63, 70, 72, 80, 84, 90, 105, 112, 120, 126, 140, 144, 168, 180, 210, 240, 252, 280, 315, 336, 360, 420, 504, 560, 630, 720, 840, 1008, 1260, 1680, 2520, and 5040 itself.

It seems that 5040 is a highly composite number (or an antiprime) by any reasonable definition, and the definition we have says that a highly composite number is a number that has more divisors than any number smaller than it. Table 2.14 lists some numbers (n) and the number $d(n)$ of their divisors, and it is immediately clear that the distribution of highly composite numbers, marked in italics in the table, is getting thinner as the numbers themselves are getting larger. This is similar to the distribution of the prime numbers.

n	1	2	3	4	5	6	7	8	9	10	11	12	24	36	48	60
$d(n)$	*1*	*2*	2	*3*	2	*4*	2	4	3	4	2	*6*	*8*	*9*	*10*	*12*

n	120	180	240	360	720	840	1260	1680	2520	5040	7560	10080
$d(n)$	*16*	*18*	*20*	*24*	*30*	*32*	*36*	*40*	*48*	*60*	*64*	*72*

Table 2.14: The First Highly Composite Numbers.

There is a simple relations between the prime factorization of a number n and the number $d(n)$ of its divisors. Take, for example, 5040. Its prime factorization is $2^4 \times 3^2 \times 5 \times 7$, which is why all numbers of the form $2^0 \cdot 3$, $2^1 \cdot 3$, $2^2 \cdot 3$, $2^3 \cdot 3$, and $2^4 \cdot 3$, are divisors of 5040, as also are $3^a \cdot 5^b$ for $a = 0, 1, 2$ and $b = 0, 1$. In general, if the prime factorization of n is $P_1^{a1} P_2^{a2} P_3^{a3} \dots P_k^{ak}$, then it has $(a1+1)(a2+2)(a3+3)\dots(ak+k)$ divisors. Thus, the prime factorization of 5040 implies that it has $(4+1)(2+1)(1+1)(1+1) = 5 \cdot 3 \cdot 2 \cdot 2 = 60$ divisors.

In early 20th century, the amazing Indian mathematician Ramanujan studied the properties of highly composite numbers and came up with three simple (necessary but not sufficient) conditions that those numbers must satisfy.

1. The prime factors of a highly composite number must be consecutive primes. Thus, if the prime factors of a number n are 2, 5, and 7 (with the prime 3 missing), then n is not a highly composite number.
2. If the prime factorization of n is $n = P_1^{a1} P_2^{a2} P_3^{a3} \dots P_k^{ak}$, then for n to be highly composite, the sequence $\{ai\}$ must be nonincreasing. Thus, the highly composite $5040 = 2^4 \times 3^2 \times 5 \times 7$, features the nonincreasing sequence 4, 2, 1, 1 of exponents.
3. The last exponent in the sequence $\{ai\}$ of exponents must be a 1 for n to be highly composite. This rule has two exceptions namely, $4 = 2^2$ and $36 = 2^2 \cdot 3^2$.

Truncatable Primes. The following prime number $L = 357686312646216567629137$, has a very special property. As we remove its digits one by one from the left, the remaining part is always a prime. Even more, there are many numbers, known as left-truncatable primes, that have this property, but L is the largest.

Such interesting numbers are easy to generate. Start with the rightmost digit. It has to be one of 2, 3, 5, and 7, because once everything to its left is removed, it has to be a prime. Now prepend another digit to its left. This must be chosen such that the resulting two-digit number will be a prime. We notice that the nine numbers 15, 25, 35, ..., 95 are all composite, as also are 12, 22, 32, ..., 92. Of the nine numbers $x3$, only 6, namely 13, 23, 43, 53, 73, and 83, are primes. Notice that digits such as 4 and 8, which cannot be the rightmost digit, may appear in other positions. The case of 7 is similar. Of the nine numbers $x7$, only 5, namely 17, 37, 47, 67, and 97, are primes. Thus, there are $6 + 5 = 11$ two-digit left-truncatable prime numbers.

Next, we prepend a third digit on the left of each of the chosen 11, and check the resulting $5 \times 9 + 6 \times 9 = 99$ numbers for primality. In this way, more and more digits are added to the left until, at 24 digits, only the above massive number L remains; the largest left-truncatable prime! The complete sequence is listed as sequence A024785 in the OEIS, reference [Sloane 08].

◇ **Exercise 2.15:** Can there be right-truncatable primes?

The Prime Constant

The first few primes are 2, 3, 5, 7, 11, one 13. Consider the unusual binary number $P = 0.r_1 r_2 r_3 \dots$ whose bits are defined by

$$r_i = \begin{cases} 1 & \text{the } i\text{th integer is a prime,} \\ 0 & \text{otherwise.} \end{cases}$$

This number, which starts with $P = 0.011010100010100010100010000\ldots_2$, is known as the prime constant (sequence A010051 in the OEIS). We can think of it as a summary of all the primes.

There are many similar binary numbers whose 1-bits and 0-bits correspond to some property (satisfied by or missing from) the integers. Such numbers are usually the elements of an infinite increasing monotonic sequence, where $a_{i+1} > a_i$ for all i. Examples are the Fibonacci sequence, the number of divisors of the integers n, the sum of those divisors, the triangular numbers, and the "look and say" sequence, Page 303. Even something such as the positions of the odd integers can be expressed with the single binary number $0.101010\ldots$.

More Types of Integers

Mathematicians, being both funny and rigorous, have defined the following types of integers:

A composite number is a positive integer that can be formed by multiplying two smaller positive integers. It can also be described as a positive integer that has at least one divisor other than 1 and itself. Every positive integer except 1 is either composite or prime.

A highly composite number is a positive integer with more divisors than any smaller positive integer has. A typical example is 24, whose seven divisors are 1, 2, 3, 4, 6, 8, 12, and 24. This is more divisors than any of the integers 1 through 23 has.

An abundant number or excessive number is a positive integer for which the sum of its proper divisors is greater than the number. The integer 12 is the first abundant number. Its proper divisors are 1, 2, 3, 4 and 6 for a total of 16.

A deficient number or defective number is a positive integer n, the sum of whose divisors is less than $2n$. Equivalently, it is a number for which the sum of proper divisors (the aliquot sum) is less than n. For example, the proper divisors of 8 are 1, 2, and 4, and their sum is less than 8, so 8 is deficient.

A primitive abundant number is an abundant number whose proper divisors are all deficient numbers.

A highly abundant number is a natural number with the property that the sum of its divisors (including itself) is greater than the sum of the divisors of any smaller natural number.

A superabundant number is a natural number n for which, for all $m < n$

$$\frac{\sigma(m)}{m} < \frac{\sigma(n)}{n},$$

where σ is the sum of all positive divisors of n, including n itself. The first few superabundant numbers are 1, 2, 4, 6, 12, 24, 36, 48, 60, and 120.

A number n is said to be colossally abundant if there is an $\epsilon > 0$ such that for all $k > 1$,

$$\frac{\sigma(n)}{n^{1+\epsilon}} < \frac{\sigma(k)}{k^{1+\epsilon}},$$

The first such numbers are 2, 6, 12, 60, 120, 360, 2520, 5040, and 55440.

A perfect number is a natural number that is equal to the sum of all its proper divisors. Such numbers are rare and searching for them requires powerful computing resources. All the perfect numbers known so far are even, and the problem of whether odd perfect numbers exist is often considered the oldest important unsolved mathematical problem.

Poem on Odd Perfect Numbers

In the realm of numbers, rare and grand,
Lies a mystery, hard to understand.
Perfect numbers, even and bright,
Yet odd ones remain out of sight.

For centuries, the search goes on,
From dawn to dusk, from dusk to dawn.
Mathematicians with minds so keen,
Ponder this puzzle, unseen, serene.

Even perfect numbers, easy to find,
Their formula simple, defined and kind.
But where's the odd one, elusive, rare?
In number theory, it's a phantom there.

Legends whisper, proofs take flight,
Yet odd perfect numbers stay out of sight.
A mathematical quest, a timeless dream,
In the world of numbers, a silent scream.

By `chatGPT 4o`, June 2024.

A semiperfect (or pseudoperfect) number is a natural number n that is equal to the sum of all or of some of its proper divisors. The first few semiperfect numbers are 6, 12, 18, 20, 24, 28, 30, 36, and 40.

A weird number is a natural number n that satisfies the following:

1. It is abundant; the sum of its proper divisors (divisors including 1 but not itself) is greater than n.

2. No subset of those divisors sums to n itself; thus, n is not semiperfect.

The smallest weird number is 70. Its proper divisors are 1, 2, 5, 7, 10, 14, and 35. Their sum is 74, but no subset of these divisors sums to 70.

A deficient (or defective) number is a positive integer n, the sum of whose divisors is less than $2n$. It can also be described as a number for which the sum of proper divisors (the so-called aliquot sum) is less than n. For example, the proper divisors of 8 are 1, 2, and 4, and their sum is less than 8, so 8 is deficient.

The following number types are not formally recognized. They may be how certain recreational mathematicians spend their leisure time. An apocalyptic power is a number of the form 2^n whose decimal representation has the string "666" somewhere. Examples are 2^{157}, 2^{192}, 2^{218}, and 2^{666}. The pair 2^{2210} and 2^{2212} are special, because they are apocalyptic powers and also 666-digits-long each, which makes them apocalypse numbers. A Goliath number is one that has 666 consecutive 6's in its decimal representation. Obviously such a number must be very large, which complicates the search for Goliaths. However, mathematicians, being hard workers and fun loving, may develop methods to define and find such large (albeit not very interesting) numbers.

◇ **Exercise 2.16:** Explain why the following relations exists:

- Many keypads are organized in three rows, three columns, and two diagonals (a total of eight) as follows

  7 8 9
  4 5 6
  1 2 3 .

  From each of these eight 3-digit numbers we can generate two 6-digit numbers as follows 123321, 321123, 456654, 654456, 789987, 987789, 741147, 147741, 852258, 258852, 963369, 369963, 753357, 357753, and 951159, 159951. Why are all 16 numbers divisible by 37?

- Why is the 12th prime number 37, while the 21st prime is 73?

- How do we explain the fact that the seven numbers 31, 331, 3,331, 33,331, 333,331, 3,333,331, and 33,333,331 are all primes?

- The figures listed in Table 2.15 were obtained by computing the first 21 powers of 2 and summing the digits of each again and again until only one digit is left in each of the 21 sums. Explain why the resulting sequence repeats every six numbers.

```
l1 = Table[Total[IntegerDigits[2^n]], {n, 0, 20}]
l2 = Table[Total[IntegerDigits[l1[[n]]]], {n, 0, 20}]
Table[Total[IntegerDigits[l2[[n]]]], {n, 0, 20}]

{1,2,4,8,7,5,10,11,13,8,7,14,19,20,22,26,25,14,19,29,31}

{Total[IntegerDigits[List]],1,2,4,8,7,5,1,2,4,8,7,5,10,2,4,8,7,5,10,11}

{Total[IntegerDigits[List]],Total[IntegerDigits[Total[IntegerDigits[List]]]],
 1,2,4,8,7,5,1,2,4,8,7,5,1,2,4,8,7,5,1}
```

Table 2.15: Sum of the Digits of Power of 2.

◇ **Exercise 2.17:** Prove that if p is a prime, then $\sqrt{p}$ is irrational.

2.4 Rationals and Irrationals

A rational number has the form n/m where n and m are integers and m is nonzero. The set of rational numbers includes the integers as the special case $m = 1$. The decimal expansion of a rational number is either finite or consists of a repeating pattern (a repetend). The repetend can be a single digit, such as $1/3 = 0.333\ldots$, or several digits, such as $1/7 = 0.142857\,142857\ldots$, but it is finite.

The proof of this property is elementary. Each step in the long division n/m results in a quotient and a remainder which is an integer in the interval $[0, m-1]$. Once the remainder produced by a division step is zero, the division terminates. Otherwise, no remainder is ever zero, and each division step results in a remainder between 1 and $m-1$. If the first remainder appears again after the first k steps, the kth step starts the same remainder pattern as the first $k-1$ steps, and the repetend's length is $k-1$ digits. There can be at most $m-1$ different remainders in the first $m-1$ steps, which is why the longest repetend is $m-1$ digits.

The converse is also true. any decimal number with a finite repetend must be rational, as the following example makes clear. Given a number of the form $A = 0.7\,162\,162\,162\ldots$, we notice that its repetend length r is 3. We first multiply it by a power of 10, such that the decimal point is in front of the first repetend

$$10A = 7.162\,162\,162\ldots,$$

then multiply by 10^3 or, in general, 10^r, to obtain

$$10,000A = 7\,162.162\,162\ldots.$$

These two decimal expansions are still infinite, so subtracting them yields $10\,000A - 10A = 9990A = 7155$, which implies that A is the rational $53/74$.

Once the basic properties of rational numbers are known, the irrational numbers are simply defined as those real numbers that do not satisfy these properties. Thus, an irrational number is one whose decimal representation is infinite and does not feature any repeating patterns. It may come as a surprise to learn (in Chapter 4) that the sets of integers and rationals have the same cardinality (the same "kind" of infinity), while the set of the irrationals is bigger in the sense that it features a bigger cardinality.

> Upon learning that no matter how far the decimal expansion of the square root of two is carried out it will never become periodic, young Julia Robinson claimed:
>
> "I didn't see how anyone could know that all they could know is that the expansion had not become periodic in the part that was calculated. I tried to check it with my newly acquired skills at extracting square roots but finally gave up. After all, I couldn't go on forever either!"
>
> —Julia Robinson.

To get a deeper understanding of the relation between integers and rationals, consider the following: Between any two consecutive integers, such as n and $n+1$, there are no other integers; the integers form a discrete sequence. The rationals, however, form a dense set. Between any two rationals, even if they are very close, there is another

rational. Given the two rationals $r_1 = a/b$ and $r_2 = c/d$, the number $r = (r_1 + r_2)/2$ lies between r_1 and r_2 and is rational because it equals the ratio $(ad + bc)/2bd$. Similarly, there are other rationals between r_1 and r, as well as between r and r_2. This implies that between any two rationals there are infinitely many other rationals.

This surprising result implies something even more baffling. Given two rationals r_1 and r_2 such that $r_1 < r_2$, there is another rational r_3 between them, such that $r_1 < r_3$. Similarly, there is another rational r_4 between r_1 and r_3 such that $r_1 < r_4$, and so on. We can construct infinitely many rationals r_5, r_6, ... that are greater than r_1 and get closer and closer to it. None of those rationals, however, will be the immediate successor of r_1 because between r_1 and any of them there will always be other rationals. We must therefore conclude that r_1 (and, in implication, any rational number) does not have any successor. There simply does not exist a number that immediately follows r_1, an unexpected, unintuitive, and baffling conclusion.

Because of these properties, the set $\mathbb{Q}$ of rational numbers is everywhere dense, but Section 4.5 shows that it has the same $\aleph_0$ cardinality as the set of integers. This is because set theory shows that it is possible to assign to every rational number an integer subscript or tag. We therefore say that the rationals are countable.

The fact that a rational number has no immediate successor means that there are no jumps in the sequence of rationals. We can move along this sequence and see the values grow continuously, without any jumps between consecutive values, because there are no consecutive rational numbers. This is hard to accept because it contradicts everything we know about the physical world. Objects in the world around us consist of discrete particles, and energy comes in discrete photons. Everything in the physical world is quantized and discrete; nothing is continuous.

When a rational number is expanded (in any number base), it becomes either a finite string of digits or an infinite repeating sequence. Thus, in decimal, 12345/1234 is the finite string 10.0040518638574, but 4/3 becomes the repeating infinite sequence 1.3333..., and 555/7 is similarly 79.285714 285714....

A fraction is reduced by dividing both sides by the same nonzero amount. Thus, dividing both 16 and 64 by 16 reduces 16/64 to 1/4. Reducing 16/64 by eliminating the 6 in the numerator and denominator is, of course, wrong, but accidentally leads to the correct result. It turns out that such cases are rare. For 2-digit numbers, this style of reduction works only for 19/95 = 1/5, 26/65 = 2/5, and 49/98 = 4/8. For bigger numbers, we mention only the case 16,666/66,664 = 1/4.

The following rationals 1/3, 2/6, 3/9, ... are equal, implying that every rational $r = a/b$ has infinitely many representations of the form na/nb for nonzero integers n. The simplest such representation, the one where a and b do not have any common factors, is called the irreducible representation of r (in our example 1/3). The irreducible form of a fraction is obtained by dividing both the numerator and denominator by their greatest common divisor. Let's select such an irreducible representation a/b at random. It seems reasonable to assume that in half the cases, the denominator b will be even, but in fact it turns out that b is even in only 1/3 of the cases. This surprising, unintuitive result is an example of beautiful mathematics.

Given two random integers p and q, we first apply Euclid's algorithm to find their greatest common divisor G. We divide $P = p/G$ and $Q = q/G$ by this divisor to obtain an irreducible fraction P/Q. It is obvious that the binary representations of P and Q

cannot both end with zeros. If $P = pp\ldots p0$ and $Q = qq\ldots q0$, then P and Q would be even and would have 2 as a common divisor. Therefore, the rational number $R = P/Q$ must have one of the following binary forms

$$\frac{pp\ldots p1}{qq\ldots q1}, \quad \frac{pp\ldots p0}{qq\ldots q1}, \quad \frac{pp\ldots p1}{qq\ldots q0}.$$

Since the original p and q were randomly selected, R can take each of these forms with equal probability. The first two forms have an odd denominator, so the probability of having an even denominator is 1/3. Elegant, simple, and surprising; in short, beautiful.

> When I considered what people generally want in calculating, I found that it always is a number.
>
> —Muḥammad ibn Mûsâ al-Khwârizmî.

We now turn to the irrationals. An irrational number is any number that cannot be written as the ratio of two integers. We intuitively feel that we "know" and can write any rational number. After all, it is easy to come up with two arbitrary integers and divide them. The irrationals, however, seem mysterious and unapproachable. If we feel or believe that a number a may be irrational, we have to prove that it cannot be written as the ratio of two integers. Such proofs are at least time consuming and may also be long, tedious, and difficult to devise and understand. As a result, only a handful of irrational numbers, such as π, $\sqrt{2}$, and e, are known. (Most roots, logarithms, and other functions are also irrational.)

As an illustration of the difficulties presented by the irrationals it is interesting to mention that it is still unknown whether combinations such as $me \pm n\pi$, for integers m and n, are rational or irrational.

Even though only a few irrationals are known, set theory (Chapter 4) shows that the cardinality of the set $\mathbb{I}$ of irrationals is greater than that of the integers and the rationals. It is possible to assign each rational number an integer tag or subscript, but it is impossible to do that with the irrationals because there simply are "too many" of them (there are infinitely many rationals and infinitely many irrationals, but the latter infinity is greater). This makes both the rationals and irrationals baffling. The rationals are both dense everywhere but are also a set of measure zero compared to their neighbors, the irrationals. The irrationals are enigmatic, because there are so many of them, yet so few are known.

Here are several examples of how to "construct" an irrational number.

- Start with a right-angle triangle of sides 1 and 2 units. The hypothenuse of this triangle must be $\sqrt{5}$ units long, because $1^2 + 2^2 = 5 = (\sqrt{5})^2$.

- Consider the number $0.12345678910|111213141516171819 20|2122\ldots$. Because of the way it is constructed (consecutive integers, starting from 1), its expansion is infinite and it does not repeat (because each integer is unique). Thus, this number—known as Champernowne constant, the concatenation of all the natural numbers—is irrational; clever! (See Page 71.)

- The fact that an irrational number has infinitely many digits after the decimal point, and they lack any repetitive patterns, can also be employed to construct irrational

numbers. We know that $\sqrt{2} = 1.414\ldots$ and $\sqrt{3} = 1.732\ldots$, so we can construct three numbers that are in the interval $(\sqrt{2}, \sqrt{3})$, that start with 1.4, 1.5, and 1.6, and that continue without repetitions. Here they are

$$1.42|43|443|4443\ldots, \quad 1.57|577|5777\ldots, \quad 1.68|69|70|71|\ldots|99|100\ldots.$$

Set theory shows that "most" real numbers must be irrational, but "most" mathematical texts claim that only a few irrational numbers—such as π, e, and ϕ—are actually known. However, it is easy to prove that the square roots of ALL integers (except perfect squares such as 4, 9, and 16) are irrational. Here is the simple proof. Assume that the integer n, which is not a perfect square, has a square root which is rational and of the form p/q where p and q are integers, the quotient p/q is in reduced form, and $q \neq 1$ (if $q = 1$ then $\sqrt{n} = p$ is an integer, implying that $n = p^2$ is a perfect square). Prime numbers are not squares, which is why the square root of a prime number is always irrational.

- The square roots of non-perfect squares are irrationals, and this can help in locating irrational numbers. What are some irrationals between 2 and 3? Since 2 equals $\sqrt{4}$ and 3 equals $\sqrt{9}$, we conclude that the square roots of 5, 6, and 7 must be irrationals, because these integers are between 4 and 9, and they are non-perfect squares. Naturally, we cannot write these numbers completely, which is why we have to use the $\sqrt{}$ symbol.

- Many mathematical constants are irrational numbers. Here are definitions of this term and some examples of constants.

  A mathematical constant is any number that is (1) well-defined and (2) significantly interesting in some way.

  A mathematical constant is an important number whose value is described by an unambiguous definition, and which is often referred to by a symbol, such as π, ϕ, $\sqrt{}$, or $\log$.

  These definitions are unsatisfactory, because the terms "well-defined," "significantly interesting," and "important" are nebulous.

  Table 2.16 lists a few constants between 2 and 3.

2.09455	Wallis Constant	2.50662	Sierpiński's constant
2.23606	Square root of 5, Gauss sum	2.62205	Lemniscate constant
2.29316	Foias constant	2.66514	Gelfond–Schneider constant
2.29558	Universal parabolic constant	2.68545	Khinchin's constant
2.37313	Lévy 2 constant	2.71828	Euler's number e
2.39996	Golden angle	2.80777	Fransén–Robinson constant
2.50290	Feigenbaum constant α	2.82641	Murata Constant

Table 2.16: Mathematical Constants in $(2, 3)$.

Any integer can be uniquely expressed as the product of prime numbers, so we can write

$$\sqrt{n} = \frac{p}{q} = \frac{p_1 p_2 \cdots p_m}{q_1 q_2 \cdots q_k},$$

where, because of the reduced form, none of p_i's equals any of the q_j's. This implies that the integer n, which equals p^2/q^2, is a fraction, and not an integer (because $q \neq 1$); and this is a contradiction to our assumption that $\sqrt{n}$ is rational.

Hence we conclude that infinitely many irrational numbers are known and can easily be specified.

The following simple arguments show why $\sqrt{2}$ is irrational. They are examples of beautiful proofs, because both are simple, easy to understand, and short. See also Page 5.

■ Assume that $\sqrt{2}$ is rational and has the form m/n for integers m and n. We divide these integers by their greatest common divisor to obtain an irreducible fraction. Thus, m and n cannot both be even because that would imply a common divisor of 2. The relation $\sqrt{2} = m/n$ can be written $m = n\sqrt{2}$ or $m^2 = 2n^2$, implying that m is even (and therefore n must be odd). Let's write $m = 2p$ and substitute $(2p)^2 = 2n^2 \Rightarrow 4p^2 = 2n^2 \Rightarrow 2p^2 = n^2$. This implies that n is even; a contradiction. Thus, our assumption that $\sqrt{2}$ is rational was wrong.

■ A geometric proof. If $\sqrt{2}$ is rational then there exist integers a and b such that $a/b = \sqrt{2}$ or

$$a^2 = 2b^2. \tag{2.1}$$

Geometrically, this means that there is an $a \times a$ square whose area equals that of two smaller $b \times b$ squares. We slide the two smaller squares into the large one, so they overlap as shown in Figure 2.17a. The total area of the two smaller squares is supposed to equal the area of the large square, which is why the area of the central overlap must equal the sum of the two small uncovered areas, Figure 2.17b. Thus, $(2b - a)^2 = 2(a - b)^2$.

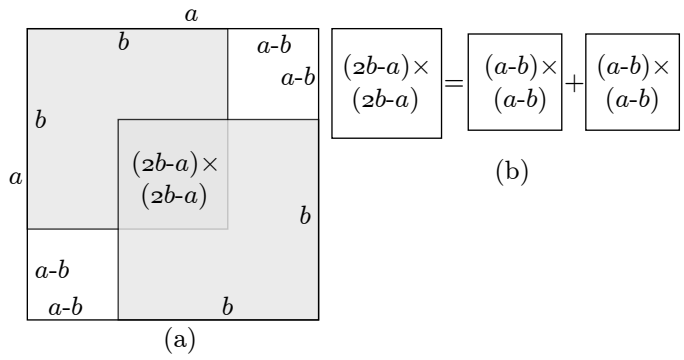


Figure 2.17: The Irrationality of $\sqrt{2}$.

This means that we have produced a new pair of integers $(2b - a)$ and $(b - a)$ that are positive (being the sides of squares) and also satisfy the relation of Equation (2.1), but are smaller than a and b. Clearly, we can continue in this way, producing an infinite list of smaller and smaller pairs of positive integers that stand in the same relation. This, however, is impossible, because integers cannot be smaller than 1, which leads to a contradiction.

See [Conway 13] for more simple proofs of the irrationality of $\sqrt{2}$.

◇ **Exercise 2.18:** (Easy). Search the mathematical literature for more irrational numbers.

The expansion of an irrational number, in any number system, is always infinite and nonrepeating, which makes them even more mysterious; there is no way to write such a number in its entirety using just digits and a decimal point.

The following discussion may shed some light on the nature of irrationals and may help the reader to come to terms with these strange numbers. Imagine a rational number $r = a/b$. We write $a = r \cdot b$, and consider a ruler of length r. When b copies of this ruler are laid end to end, they cover a distance of a units. Now imagine an irrational number t. The definition of irrationals implies that there is no ruler of length t such that laying it end to end b times (where b is any integer) will cover a units (where a is an integer). Weird!

Imagine two line segments of lengths a and b, where the ratio a/b is irrational. Such line segments are referred to as incommensurable, a term that means they have no common measure.

Here is an example of this weird behavior. The irrational number $\sqrt{2}$ equals approximately 1.414, a number that can be written as the ratio of two integers 1414/1000. If we compute ten more digits of $\sqrt{2}$, we end up with the approximate value 1.4142135623731, a number that can be written as the ratio 14142135623731/10000000000000. We can continue in this way to compute and write better approximations of $\sqrt{2}$, but since this number is irrational, its expansion is endless and nonrepeating, so there is no way to represent it *precisely* as a ratio. Such "behavior" may repulse some people, but others considers it interesting, mysterious, and beautiful. It certainly is very unintuitive.

◇ **Exercise 2.19:** An irrational number contains infinitely many digits with no repeating patterns. Can we therefore claim that such a number contains every conceivable pattern? Does π, for instance, contain your birthdate or that best-seller book you plan to write?

◇ **Exercise 2.20:** Can a sequence or a series of rational numbers converge to an irrational number? The concept itself seems irrational.

◇ **Exercise 2.21:** Show that there exist irrationals x and y such that x^y is rational.

◇ **Exercise 2.22:** Prove that $\tan 1^\circ$ is irrational.

2.4.1 Square roots and irrationality

Many (perhaps most) square roots of real numbers are irrational, but it is easy to show that the square roots of the integers are irrational except for those integers that are perfect squares. This simple, elegant proof, due to [Flanders 99], is based on an ingenious idea of Theodor Estermann's.

Given a positive integer m that is not a perfect square, we find the integer n which satisfies $n < \sqrt{m} < n + 1$, and denote $a \stackrel{\text{def}}{=} \sqrt{m} - n$, which implies $0 < a < 1$. We prove that a is irrational by way of a contradiction. Suppose that $a = p/q$, where p and

q are positive integers, $0 < p < q$, and q is already as small as possible (the latter is Estermann's key idea). Then we have

$$\frac{p}{q} = \frac{1}{\sqrt{m}-n} = \frac{1}{(\sqrt{m}-n)}\frac{\sqrt{m}+n}{(\sqrt{m}+n)} = \frac{\sqrt{m}+n}{m-n^2} = \frac{a+2n}{m-n^2}.$$

We isolate a to obtain

$$a = \frac{(m-n^2)q}{p} - 2n = \frac{(m-n^2)q-2np}{p},$$

but this represents a as a fraction whose denominator is smaller than q, which creates a contradiction.

Figure 2.18 illustrates graphically the difference between rational and irrational numbers. The red lines correspond to rational numbers and go through one or several grid points. The three lines marked π, e, and ϕ correspond to well-known irrational numbers and never pass through grid points, even though they may pass close to some points, which then become good approximations to the irrationals.

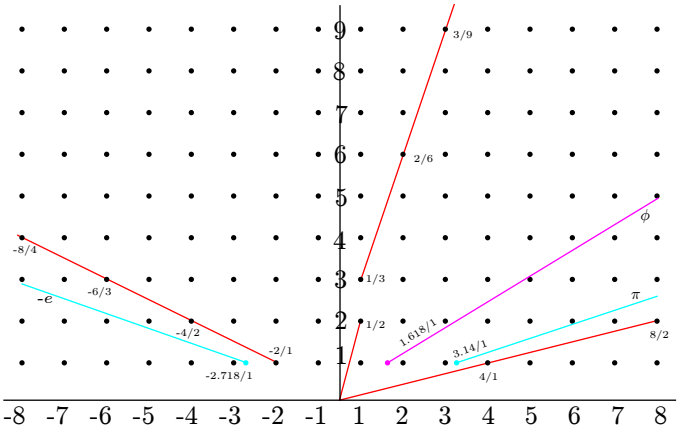


Figure 2.18: Rational and Irrational Numbers.

◇ **Exercise 2.23:** Show that the sum $\sqrt{2}+\pi$ of two irrationals is also irrational.

2.5 Continued Fractions

A continued fraction (more generally, a generalized continued fraction) is a mathematical expression of the form

$$x = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cfrac{a_4}{b_4 + \cfrac{a_5}{\ddots}}}}}. \tag{2.2}$$

This is a fraction with a denominator that is a sum that contains another continued fraction. If a continued fraction is finite, it terminates with a simple fraction. The a_i terms of a continued fraction are referred to as partial numerators, and its b_i terms are the partial denominators. The leading term b_0 is the integer part of the continued fraction.

A simple continued fraction is the important special case where all the partial numerators are ones

$$x = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \cfrac{1}{b_4 + \cfrac{1}{\ddots}}}}}. \tag{2.3}$$

Some texts use the notation a_i's for the partial denominators of a simple continued fraction.

Continued fractions are a typesetter's nightmare, which is why a compact notation is often used to describe them. A generic simple continued fraction, such as Equation (2.3), can be written either as $[b_0, b_1, b_2, b_3, \ldots]$ or as $[b_0; b_1, b_2, b_3, \ldots]$. The compact form of the generalized continued fraction of Equation (2.2) is

$$x = b_0 + \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+\cdots}.$$

An important feature of continued fractions is their ability to distinguish between rational and irrational numbers. The decimal representation of an irrational number is infinite and non-repeating, while the decimal representation of a rational number is normally finite, but sometimes is infinite and repeating. On the other hand, the continued fraction of a rational number is always finite, while that of an irrational number is always infinite.

Any irrational number x can be written uniquely as an infinite simple continued fraction $[a_0; a_1, a_2, \ldots]$, where the partial denominators a_i are integers that share the sign of x, i.e., they are all positive or all negative, except that a_0 may be zero. However, the generalized continued fraction of an irrational may not be unique. See, for example, Equations (2.6) and (2.9) for an example in the case of π.

The simple continued fraction expansion of a rational number is finite and always has the two similar forms $[a_0; a_1, \ldots a_{n-1}, a_n]$ and $[a_0; a_1, \ldots a_{n-1}, (a_n - 1), 1]$. In practice, the first, shorter expansion is used and is referred to as the canonical continued fraction expansion (or representation) of the rational number.

The simple continued fraction expansion of φ, the golden ratio, is the repeating $[1;1,1,1,\ldots]$ and that of $\sqrt{2}$ is $[1;2,2,2,\ldots]$. It turns out that all the quadratic irrationals have repeating continued fraction expansions. This is in stark contrast with the decimal representation of the same quadratic irrationals, which are random. (The precise definition of quadratic irrationals is outside the scope of this book, but it is pointed out that the square roots of all positive integers that are nor perfect squares are quadratic irrationals.)

A noble number ν is an irrational number having a continued fraction that becomes an infinite sequence of 1's at some point. Thus, $\nu = [0;a_1,a_2,...,a_n,\bar{1}]$. See [Schroeder 91] for more information. Because of their definition, noble numbers, of which φ is one, are very irrational as measured by the criterion discussed on Page 89.

Here are two techniques that illustrate how a continued fraction can be constructed recursively. The first technique starts with the simple quadratic equation $x^2-bx-1=0$, which is then rewritten as

$$x = b + \frac{1}{b+\frac{1}{x}}. \tag{2.4}$$

Replacing the x with the right-hand side of Equation (2.4), and repeating several times produces the continued fraction

$$x = b + \cfrac{1}{b+\cfrac{1}{b+\cfrac{1}{b+\cfrac{1}{b+\ddots}}}}.$$

This technique also demonstrates that quadratic equations are intimately related to continued fractions. Here is another example. Given the equation $x^2-4x-12=0$, whose solution is the integer 6, we can divide it by x (which cannot be zero) and rearrange to obtain

$$x = 4 + \frac{12}{x}.$$

We now replace the x by the entire right-hand side to end up with the start of a continued fraction

$$x = 4 + \frac{12}{4+\frac{12}{x}}, \tag{2.5}$$

and can continue in this way forever. Notice that this type of continued fraction can be infinite even though it converges to the rational 6. (The continued fraction of an integer is the integer itself, but Equation (2.5) includes an unknown and is therefore an equation whose solution is 6, and not the simple continued fraction expansion of the number 6.) Also, this continued fraction is not simple because its numerators are not all 1's.

This is not a practical way to solve the original quadratic equation, but it can provide a way to compute a sequence of better and better approximate solutions. In our example, the first few approximations are

$$x_0 = 4, \quad x_1 = 4 + \frac{12}{4} = 7, \quad x_2 = 4 + \frac{12}{4+\frac{12}{4}} = 4 + \frac{12}{7} = \frac{40}{7} = 5.7,$$

and they oscillate about the final value of 6, while getting closer to it in each iteration.

The second technique starts with the well-known irrational number π. We know that it equals 3.1415 ..., so we first write it as $3 + (\text{a bit})$, and then try to refine this using fractions instead of simply appending digits to its decimal representation. The "bit" mentioned earlier must be less than 1, so we try to write it as a fraction $3 + (1/x)$. Using a calculator or mathematical software we solve the equation `Solve[Pi == 3 + 1./x, x]` and discover that $x = 7.06251$, a little more than 7. Now we can write $\pi = 3 + 1/(7 + y)$, and with the help of `Solve[Pi == 3 + 1./(7+y), y]` we obtain $y = 0.0625133$. Continuing in this way, writing $\pi = 3 + 1/(7 + 0.0625133)$ and looking for z that satisfies $1/z = 0.0625133$, we find that $z = 15.9966$, so we write

$$\pi \approx 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{\cdots}}}.$$

After finding the solution $w = 0.0625133$ to $1/w = 15.9966$, we proceed a few more steps in this way to obtain

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{\ddots}}}}}. \tag{2.6}$$

Figure 2.19 lists the first 100 coefficients in the continued fraction expansion of π.

```
ContinuedFraction[Pi, 100]
3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, \
84, 2, 1, 1, 15, 3, 13, 1, 4, 2, 6, 6, 99, 1, 2, 2, 6, 3, 5, 1, 1, 6, \
8, 1, 7, 1, 2, 3, 7, 1, 2, 1, 1, 12, 1, 1, 1, 3, 1, 1, 8, 1, 1, 2, 1, \
6, 1, 1, 5, 2, 2, 3, 1, 2, 4, 4, 16, 1, 161, 45, 1, 22, 1, 2, 2, 1, \
4, 1, 2, 24, 1, 2, 1, 3, 1, 2, 1, 1, 10
```

Figure 2.19: First 100 Coefficients of the Continued Fraction of π.

The continued fraction of a rational number is finite and may even be very short, but it is non-unique and may end in one of two different ways. Consider the rational 1386/50 whose continued fraction expansion is especially short

$$\frac{1386}{50} = 27 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \frac{1}{3}}}}}.$$

It is easy to verify that this is correct by starting from the bottom and proceeding in steps as follows:

$$1 + \frac{1}{3} = \frac{4}{3}; \quad 1 + \frac{3}{4} = \frac{7}{4}; \quad 2 + \frac{4}{7} = \frac{18}{7}; \quad 1 + \frac{7}{18} = \frac{25}{18}; \quad 27 + \frac{18}{25} = \frac{693}{25},$$

and $693/25 = 1386/50 = 27.72$.

The continued fraction of an irrational number, however, is infinite, which makes it impossible to fully verify it. The best that can be done is to chop it level by level, compute the part that remains, and see how fast the partial values seem to converge to a final value.

Equation (2.6) implies that the first few partial values (also known as convergents) of the continued fraction of π are

$$3 + 1/7 = 21/7 + 1/7 = 22/7,$$

$$7 + \frac{1}{15} = \frac{106}{15} \to 3 + \frac{15}{106} = \frac{333}{106}, \text{ and}$$

$$15 + \frac{1}{1} = 16, \to 7 + \frac{1}{16} = \frac{113}{16}, \to 3 + \frac{16}{113} = \frac{355}{113}.$$

The recursive recurrence formulas for continued fractions, originated by Pietro Cataldi and rewritten in modern form by Euler and by John Wallis, can simplify the process of computing the partial values. Employing the notation of Equation (2.2), they start with

$$A_{-1} = 1,\ A_0 = b_0,\ A_{n+1} = b_{n+1}A_n + a_{n+1}A_{n-1},$$
$$B_{-1} = 0,\ B_0 = 1,\ B_{n+1} = b_{n+1}B_n + a_{n+1}B_{n-1},$$

and then compute the partial values as A_n/B_n.

Applying these recurrence formulas to the continued fraction expansion of φ, the golden ratio, it is clear that successive values of A_n and B_n, which are $(1, 2, 3, 5, 8, 13\ldots)$ and $(1, 1, 2, 3, 5, 8, 13\ldots)$, respectively, are the Fibonacci numbers. The successive convergents are therefore the ratios 1/1, 2/1, 3/2, 5/3, 8/5, ... of consecutive Fibonaccis, and it well known that those ratios converge to φ.

Table 2.20 shows that the partial values of the continued fraction of π converge quickly, while those of the continued fraction of ϕ converge very slowly.

π	3.1415926535	ϕ	1.6180339887
3	3.0	1	1.0000000000
$\frac{22}{7}$	3.14 285714285714	2	2.0000000000
$\frac{333}{106}$	3.1415 0943396226	$\frac{3}{2}$	1.5000000000
$\frac{355}{113}$	3.141592 92035398	$\frac{5}{3}$	1.6666666666
$\frac{103993}{33102}$	3.141592653 01190	$\frac{8}{5}$	1.6000000000

Table 2.20: Converges of the Continued Fractions of π and ϕ.

The continued fraction expansion of a rational number is always finite. This is because the continued fraction expansion of a rational number r can be obtained by repeatedly applying the Greatest Common Divisor (Euclidean) algorithm to r and 1. The Euclidean algorithm terminates after a finite number of steps, so the continued fraction expansion must also be finite. On the other hand, the continued fraction expansion

of an irrational number is always infinite, because repeatedly applying the Euclidean algorithm to it and to 1 never terminates.

The continued fraction (2.6) cannot be fully written because it is infinite. Adding more steps improves the approximate result that it produces, but the important point is the last step, $1/(292+u)$, listed in (2.6). The first three steps add fractions that are a bit larger than 1/7, 1/15, and 1/1. Those are large fractions, but step 4 adds the much smaller fraction $1/(292+u)$, and therefore constitutes a very little improvement to the approximation of π (u, of course, is less than 1). Future steps may add larger fractions, but this small contribution means that the continued fraction approximation of π given by the previous steps was already good. Thus, π, which is irrational, can be approximated very accurately by only three levels of continued fractions. Those three levels (in fact, any finite number of levels) describe a rational number, which is why we can confidently claim that, in the sense discussed here, π is not far from being rational, and is therefore not a very irrational number!

Insert: Continued fractions and the concept of irrationality

How can we measure the irrationality of an irrational number? That elusive attribute that makes it either quick and easy or very slow and tedious to compute a close rational approximation of a given irrational number. After all, given an irrational number a, there are infinitely many rational numbers that get closer and closer to a. As an irrational, a has infinitely many digits after the decimal point, and if we cut off this infinite sequence at some point, we are left with a rational that may be close to a. This suggests the following measure of irrationality. Locate the first point where chopping this infinite sequence of digits would leave a rational that is a good approximation of the original irrational a, but chopping it at a later point would not significantly improve that approximation.

As an example of such chopping, consider an irrational number $a = 0.3540999325\ldots$. We replace the part `0999` with `1000` to obtain a number $b = 0.3541000325\ldots$ that is only slightly greater than a. We know that anything following the `325` will not change the value of a significantly, which is why we can reasonably claim that the irrationality of a is in the order of three digits. With just three digits we obtain a rational that approximates a, but adding more digits would not improve that approximation.

On the other hand, an irrational number of the form $0.555\ldots5556555\ldots$ is more irrational than a, because none of its less-significant digits is significantly smaller than the first digits. The question now becomes what is the most irrational number according to our criterion? It should be a number of the form $ddd\ldots$, but such a number is rational. This is why using the actual digits of a number to determine its irrationality is not a good choice.

A deeper understanding of irrational numbers is obtained when we try to estimate the irrationality of an irrational number not by its digits, but by means of the coefficients of its continued fraction representation. The excellent video at [mostIrrat 23] is recommended for this topic.

The earlier claim that π is not a very irrational number is an unexpected result. We intuitively expect π to be a peculiar, unusual number, especially since it is so universal,

has many applications in mathematics, engineering, and science, and shows up in so many unanticipated places.

An interesting conclusion that can be drawn from this experience with π is that the more irrational a number is, the more large fractions its continued fraction expansion features and the further down the first small fraction will be found. The most extreme case is the most irrational number, the one whose simple continued fraction expansion contains only 1's.

Trusting to our mathematical intuition, we try the well-known and important irrational number $\sqrt{2}$ and discover that it really is a bona fide irrational because its simple continued fraction approximation is

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{\ddots}}}}}.$$

Each step improves the approximation by adding a fraction that is slightly larger than $1/2$, thereby declaring $\sqrt{2}$ to be quite good as an irrational. Notice also that this continued fraction is not just simple (its numerators are all ones), it is also very simple because its denominators are all 2.

◇ **Exercise 2.24:** Show how to derive the continued fraction expression of $\sqrt{2}$.

Surprisingly, it is easy to prove that the most irrational number, the one whose continued fraction approximation contains only 1's, is φ (or ϕ), the golden ratio, Subsection 2.11.4. We start with the infinite continued fraction that contains only 1's, i.e., the one where every step adds a fraction that is the largest possible and is slightly greater than $1/2$,

$$x = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}}}, \tag{2.7}$$

and, being infinite, we simply rewrite it as $x = 1 + 1/x$, which is the quadratic equation $x^2 - x - 1 = 0$, that we solve by completing the square $(x - 1/2)^2 = 5/4$, to obtain $x = (1 \pm \sqrt{5})/2$. The "+" sign result in φ while the "−" yields $-1/\varphi$. The discussion of sunflower seeds on Page 295 uses this unique property of φ.

And now, what is your opinion? Is φ important because it is the most irrational number or because it is the ratio that is the most pleasing to the eye? You, gentle reader, be the judge.

◇ **Exercise 2.25:** What is the continued fraction of e?

◇ **Exercise 2.26:** Continued fractions are useful, but can also lead to confusion as the following exercise demonstrates. Start with the simple identity

$$1 = \frac{2}{3 - 1}$$

and substitute this relation for the 1 in the denominator

$$1 = \frac{2}{3 - \frac{2}{3-1}}.$$

This can be the start of the infinite continued fraction

$$1 = \cfrac{2}{3 - \cfrac{2}{3 - \cfrac{2}{3 - \cfrac{2}{\ddots}}}}.$$

However, it is also true that

$$2 = \frac{2}{3 - 2},$$

which is why similar substitutions produce the following continued fraction

$$2 = \cfrac{2}{3 - \cfrac{2}{3 - \cfrac{2}{3 - \cfrac{2}{\ddots}}}}.$$

The surprise is that the right-hand sides of these continued fractions are identical, suggesting that either $1 = 2$ or that there is a mistake somewhere. Your task, if you still believe that $1 \neq 2$, is to explain this seeming paradox.

Continued Fractions and Quadratic Equations

The general quadratic equation $ax^2 + bx + c = 0$ is often solved by the well-known formula which is obtained by completing the square. However, this formula contains a square root, which is why it is sometimes preferable to solve such an equation by means of a continued fraction. The idea is to manipulate the quadratic equation in two steps as follows

$$x^2 + bx = -c, \quad x + b = \frac{-c}{x}, \quad x = -b - \frac{c}{x},$$

and then apply the last result to itself recursively by substituting it for the x. The result is the continued fraction

$$x = -b - \cfrac{c}{-b - \cfrac{c}{-b - \cfrac{c}{-b - \cfrac{c}{\ddots}}}}. \tag{2.8}$$

However, a quadratic equation must have two solutions (which may be identical) and the continued fraction, if it converges at all, may often converge to one value only. Also, a quadratic equation may have complex solutions, but Equation (2.8) can produce

only real values. The following rules show how to determine if a continued fraction can serve as the solution of a given quadratic equation:

If $b = 0$, the continued fraction solution diverges; its convergents alternate between 0 and ∞. If $b \neq 0$, there are three cases.

1. If the discriminant $b - 4ac$ is negative, the continued fraction cannot produce a solution. If the continued fraction is constructed and evaluated, its convergents wander around in either a regular or a chaotic way, never approaching a limit; we say that it diverges by oscillation.

2. If the discriminant is zero, the two solutions are identical, and are produced by the continued fraction.

3. If the discriminant is positive, the equation has two distinct real roots, and the continued fraction converges to the larger (in absolute value) of these. The reader may find it intriguing to note that the rate of convergence depends on the absolute value of the ratio between the two roots; the farther that ratio is from unity, the faster the continued fraction converges.

Euler's continued fraction formula

One of the many achievements of Leonhard Euler is a formula relating the sum of partial products to a continued fraction as follows

$$a_1 + a_1a_2 + a_1a_2a_3 + a_1a_2a_3a_4 = \cfrac{a_1}{1 - \cfrac{a_2}{1+a_2 - \cfrac{a_3}{1+a_3 - \frac{a_4}{1+a_4}}}}.$$

This formula speeds up the computations of many functions as the following example demonstrates. The inverse tangent function is traditionally represented as the infinite sum

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots,$$

but once this is converted to the sum of partial products

$$x + x\left(\frac{-x^2}{3}\right) + x\left(\frac{-x^2}{3}\right)\left(\frac{-x^2}{5}\right) + x\left(\frac{-x^2}{3}\right)\left(\frac{-x^2}{5}\right)\left(\frac{-x^2}{7}\right) + \cdots,$$

it can be converted to a continued fraction by means of Euler's formula as follows

$$\cfrac{x}{1 - \cfrac{\frac{-x^2}{3}}{1+\frac{-x^2}{3} - \cfrac{\frac{-3x^2}{5}}{1+\frac{-3x^2}{5} - \cfrac{\frac{-5x^2}{7}}{1+\frac{-5x^2}{7} - \ddots}}}}.$$

Having obtaining this result, we apply it to constructing a continued fraction for π by observing that $\tan^{-1}(1) = \pi/4$. When this is substituted in the previous result, it

yields the elegant generalized continued fraction for π

$$\pi = \cfrac{4}{1+\cfrac{1^2}{2+\cfrac{3^2}{2+\cfrac{5^2}{2+\cfrac{7^2}{2+\cfrac{}{2+\cdots}}}}}}. \tag{2.9}$$

Note. Equations (2.6) and (2.9) are different continued fraction expansions of π. The simple continued fraction expansion of an irrational is unique, but Equation (2.6) is a generalized expansion.

One practical application of continued fractions is their ability to provide good rational approximations of irrational numbers. It has long been known that 22/7 and 355/113 are excellent approximations of π, and it is now obvious how these rationals are found. They are the results of chopping the continued fraction of π after several steps. Table 2.21 lists the first seven rational approximations of π obtained from its continued fraction.

3
22/7
333/106
355/113
103993/33102
104384/33215
208341/66317

Table 2.21: Seven Rational Approximations of π.

From the earlier discussion of φ and its irrationality we know that the earlier a large number appears in the continued fraction of an irrational number, the better rational approximation is obtained when the continued fraction is chopped off at that point. Again using π as an example, we notice that the obvious rational approximation 3.14 or 314/100 is very similar to $333/106 = 3.14151$, yet simple calculations indicate that the differences between these values and π are large. They are $\pi - 3.14151 = 0.0000826536$ and $\pi - 3.14 = 0.00159265$ and the ratio of these differences is 19.269. Thus 333/106 is more than 19 times a better approximation of π than 3.14. Similar checks prove that the next approximation, 103993/33102, is 12712 times better than 3.1416.

An extreme example of a very large number appearing early in the continued fraction of an irrational number is 16539, which appears in the 6th place is the expansion of π^4. The Mathematica statement `ContinuedFraction[Pi^4, 12]` produces the sequence $[97, 2, 2, 3, 1, 16539, 1, 6, 7, 6, 8, 6]$, and the rational approximation to π^4 obtained when the expansion is chopped off at 16539 is 2143/22. This is why the difference between π and the number $(2143/22)^{1/4} \approx 3.1415926525826463$ is about 10^{-9}.

In contrast, the first large number, comparable to 16539, in the continued fraction expansion of π, is 20776, which appears in the 432nd position, very late to be of use in obtaining a rational approximation to this important constant.

It is now also clear that it is difficult to find a satisfactory rational approximation to φ, because all the denominators in its continued fraction expansion are 1's. There is no large denominator.

Gears and their ratios

A gear or gearwheel is a rotating machine part typically used to transmit rotational motion and/or torque by means of a series of teeth that engage with compatible teeth of another gear or other part. (Quoted from Wikipedia.)

The large class of toothed wheels encompasses several types such as cogs, sprockets, pulleys, and gears, each of which consists of an entire family of designs and shapes.

Our interest in gears stems from the fact that meshing gears involves ratios of numbers of teeth. In a car, for example, if we want one revolution of the wheels to correspond to five revolutions of the engine, we need two toothed wheels whose numbers of teeth are in the ration of 1:5.

The most common gear ratios used in cars are as follows: In 1st gear 1:3.5 up to 1:4.5. In 2nd gear 1:2 up to 1:2.5. In 3rd gear 1:1.5 up to 1:1.8. In 4th gear 1:1 (straight connection) up to 1:1.3. In 5th gear 1:0.8 (wheels spin faster than engine) up to 1:1.

As long as a gear ration is rational, it is easy to decide on the number of teeth needed. A ratio of 1:1.8 can be represented as 10:18, 50:90, 70:126 and others. Gears with just 10 and 18 teeth may not mesh properly, so the actual numbers of teeth depend on the sizes of the toothed wheels, which in turn depend on the amount of power to be sent from the engine to the car wheels. Gears in a big truck must be bigger than those in a small passenger car, and must therefore contain a larger number of teeth.

What if a gear ratio should be an irrational number? Such a number by definition cannot be represented as the ratio of two integers, which is why the designer of such a gear system must find a proper rational approximation of the irrational in question. The approximation should be as accurate as possible, but should not contain too many or too few teeth.

As an example, consider the gear ratio 1: $\sqrt{2}$. This is a bigger toothed wheel driving a smaller wheel, such that each revolution of the big driving wheel turns the smaller wheel about 1.414 turns. The first few continued fraction approximations of $\sqrt{2}$ are 3/2, 7/5, 17/12, 41/29, and 99/70. The rational 99/70, or any of its multiples, is so close to the actual $\sqrt{2}$, that the difference between them is -0.0000721519 or about 0.0072%. Thus, it would take 13859.6 revolutions of the big wheel to cause an accumulated error of one revolution in the small wheel.

Why would one want an irrational gear ratio? Perhaps the best example is an orrery, a mechanical or electro-mechanical model of the Solar System that shows the relative positions and motions of the planets and moons at various times according to the heliocentric or geocentric models. The first modern orrery was made around 1712 by John Rowley who named it for his patron Charles Boyle, 4th Earl of Orrery.

Consider the (unusual) orrery that features the Earth together with Pluto and its moon Charon. The mean time it takes Charon to orbit around Pluto is 6.387 (Earth) days. This is not an irrational number, but will serve as an example. The problem is to compute the continued fraction expansion of 6.387 and its convergents, and this is easily done with mathematical software, such as Mathematica. The two statements `ContinuedFraction[6.387, 100]` and `Convergents[6.387, 5]` produce the expan-

The antikythera mechanism, discovered in 1901 in a wreck off the Greek island of Antikythera in the Mediterranean Sea, exhibited the diurnal motions of the Sun, Moon, and the five planets known to the ancient Greeks. It has been dated between 205 to 87 BC. The mechanism is considered one of the first orreries. It was geocentric and used as a mechanical calculator to calculate astronomical positions.

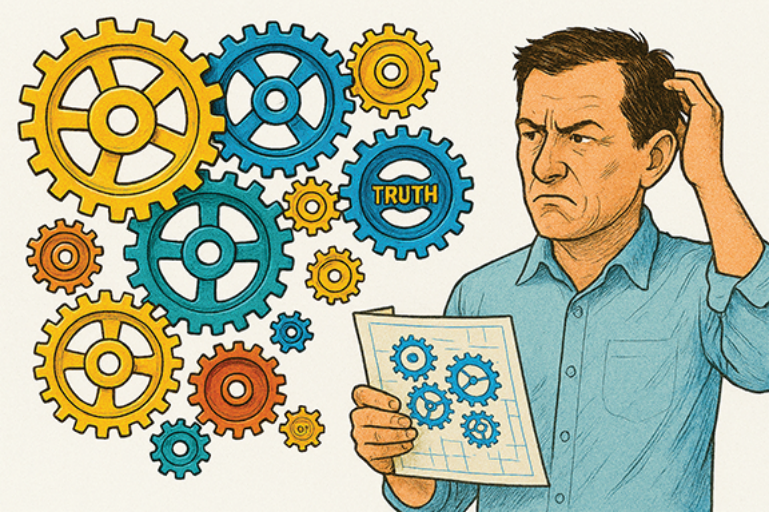


Figure 2.22: Image by MS Copilot, Sept. 2025.

sion $(6, 2, 1, 1, 2, 2, 10)$ and the list of convergents $(6, 13/2, 19/3, 32/5, 83/13)$. The most accurate convergent is $83/13$ and it, or its double $166/26$, involve numbers that are reasonable for manufacturing toothed wheels (such numbers should ideally be between 10 and 200 teeth).

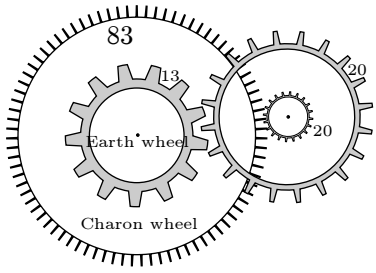


Figure 2.23: Toothed Wheels for a 6.387 Gear Ratio.

Figure 2.23 Illustrates a possible arrangement of cogs (the simplest toothed wheels)

to implement this gear ratio. The two wheels on the left are concentric but rotate at different rates. For every 6.387 rotations of the small wheel (Earth), the large wheel (Charon) rotates once. The two wheels on the right rotate together. Their function is to synchronize the rotations of the left-hand side wheels.

Continued Fractions and Probabilities

There is an unexpected relation between continued fractions and probability. The beginning of this section includes the short paragraph:

> Any irrational number x can be written uniquely as an infinite simple continued fraction $a_0, a_1, a_2, \ldots$, where the partial denominators a_i are integers that share the sign of x, i.e., they are all positive or all negative, except that a_0 may be zero. The continued fraction expansion of a rational x is finite and is not always unique.

This paragraph is now turned on its head to conclude that every infinite sequence of same-sign integers is the continued fraction expansion of a unique irrational number.

The space of infinite sequences of same-sign integers is countably infinite, and the distribution of numbers in this space is uniform. All integers, such as 1, 292, and 16539, appear with the same probability. The surprise is that there is a nonuniform distribution that holds true for almost every irrational number. In this distribution, the probability of the nonzero integer k is

$$P(k) = \frac{\ln\left[1 + \frac{1}{k(k+2)}\right]}{\ln 2}, \tag{2.10}$$

and it goes down to zero very quickly as k gets bigger, because the limit of Equation (2.10) is $1/k^2$.

The Mathematica statements

```
Pr[k_] := Log[E, 1.+(1/(k(k+2)))]/Log[E,2]
Table[Pr[k], {k, 1, 9}]
```

produce $(0.415, 0.167, 0.093, 0.059, 0.041, 0.030, 0.0233, 0.018, 0.015)$. Similarly, the probability of the rare denominator 292, which appears early in the expansion of π is $P(292) = 0.0000168051$ or 0.00168%, and the probability of 16,539 is the incredibly small $5.27356 \cdot 10^{-9}$.

A related question is the average of the infinite sequence a_i of the continued fraction expansion of those irrational numbers that obey this unusual distribution. Their arithmetic average is the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} a_k$$

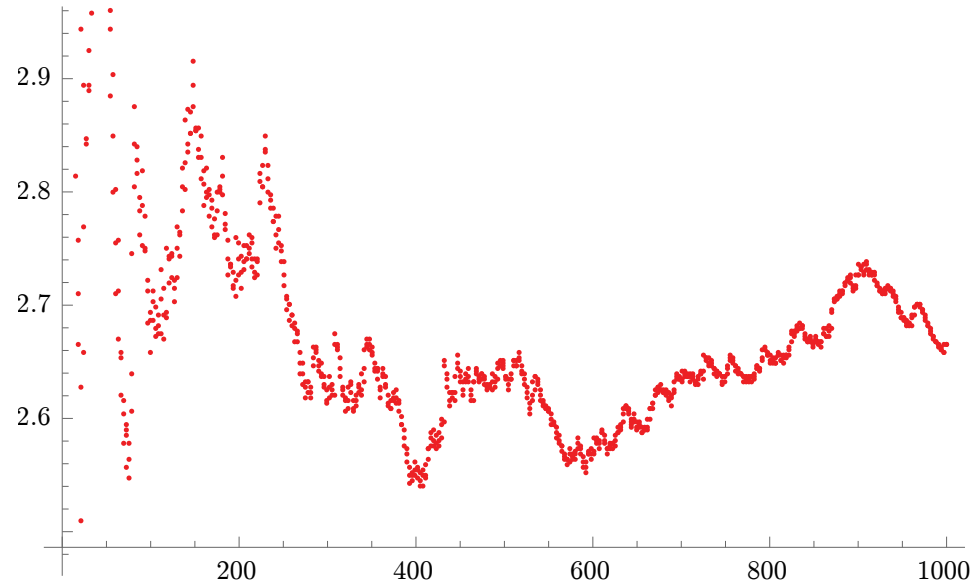
which is of no use because it diverges, but the geometric mean tends to a limit known as Khinchin's constant K_0 (Page 62). Thus,

$$\lim_{n\to\infty} (a_1 a_2 a_3 \ldots a_n)^{1/n} = 2.6854520010\ldots.$$

The fact that this constant is a small number implies that the elements (denominators) a_i in the infinite simple continued fraction expansion of almost all irrational numbers are

small. There are exceptions to this behavior, as for example, e. The continued fraction expansion of this number contains infinitely many 1's as well as all the even integers, so its geometric mean grows without limit, albeit very slowly.

Figure 2.24 illustrates how the first 1000 denominators in the expansion of π start converging toward K_0.




```
L = ContinuedFraction[Pi, 1000];
M = {};
Do[AppendTo[M, GeometricMean[L[[1 ;; i]]]], {i, 1000}];
ListPlot[M, PlotStyle -> {Red}]
```

Figure 2.24: The Geometric Means of the First 1000 Denominators in the Expansion of π.

> In short, I never yet encountered the mere mathematician who could be trusted out of equal roots, or one who did not clandestinely hold it as a point of his faith that x squared + px was absolutely and unconditionally equal to q. Say to one of these gentlemen, by way of experiment, if you please, that you believe occasions may occur where x squared + px is not altogether equal to q, and, having made him understand what you mean, get out of his reach as speedily as convenient, for, beyond doubt, he will endeavor to knock you down.
>
> —Edgar Allan Poe, *The purloined letter.*

2.6 Real Numbers

We believe we have an intuitive grasp on integers due to our ability to count cows, but when considering real numbers, their unexpected properties and behavior render them impossible to truly understand.

The set $\mathbb{R}$ of real numbers consists of the intermingled sets of the rationals $\mathbb{Q}$ and irrationals $\mathbb{I}$. The ancient Greeks already knew that $\sqrt{2}$ is irrational. The real numbers can also be divided into algebraic and transcendental numbers. The former is the set of all the reals that are solutions of algebraic equations with rational coefficients.

We are familiar with many integers (0, 1, 7, 10, and 2^8 immediately come to mind), but we know (and use) very few irrational numbers ($\sqrt{2}$, e, and π are the main examples). This is why people intuitively feel that most real numbers must be rational and the irrationals constitute a small minority. Similarly, it is easy to believe that most reals are algebraic and the transcendental numbers are rare. However, set theory, the creation, in the 1870s, of Georg Cantor, shows that there are different kinds of infinities, that the reals constitute a greater infinity than the integers (the integers are said to be countable, while the reals are not), that the rational numbers are countable, while the irrationals are uncountable, and similarly, that the algebraic numbers are countable, while the transcendentals are uncountable; completely counterintuitive notions.

Today, we believe in the existence of atoms. If we start with a chunk of matter, cut it into pieces, cut each piece into smaller pieces, and repeat this process a finite number of times, we will eventually arrive at individual atoms or even their constituents. The real numbers, however, are very different. They can be represented as points along an infinitely long number line, but they are everywhere dense on this line. Any segment on the number line, as short as we can imagine, contains an (uncountable) infinity of real numbers. By repeatedly segmenting and producing shorter and shorter segments a finite number of times, we cannot arrive at a segment containing just one number.

Imagine a straight segment of unit length and denote its midpoint by M. Now assign to each point Q on the segment a coordinate d_Q that is its distance from M. Points to the right of M have positive coordinates up to $1/2$, while points to the left of M have negative coordinates down to $-1/2$. Now imagine an infinitely-long straight line. It does not have a midpoint, so we select a point N at random as our midpoint. Each point P on the line is assigned as its coordinate its distance d_P from N. Clearly, these distances vary from $-\infty$ to $+\infty$. We now show that our short segment and long line have the same number of points (the same infinity) by associating a point Q on the segment for each point P on the line. We do this with the simple function $d_Q = \arctan(d_P)/\pi$.

For those who forgot their school trigonometry, the arctan function is the inverse of tan (or tangent). This is a multi-valued function whose main branch varies from $\arctan(-\infty) = -\pi/2$ to $\arctan(0) = 0$ to $\arctan(\infty) = \pi/2$. Thus $\arctan(d_p)/\pi$ varies from $-1/2$ to $+1/2$ and covers the coordinates d_Q of all the points Q on the 1-unit-long segment.

This simple function associates a point Q on the segment with every point P on the infinite line, thereby showing that the segment and the line have the same infinity of points, an unintuitive result. This result is true for any segment, regardless of its length, which implies that any interval of real numbers, even the shortest imaginable, contains the same infinity of reals as the interval $[-\infty, +\infty]$.

Even more surprises await the reader. It is easy to show that the (infinite) number of two-dimensional points equals the (infinite) number of reals. All we need is a function that maps each two-dimensional point (x, y) to a real number R, such that different points will map to different reals and every point will map to some real number.

A two-dimensional point is a pair of coordinates (real numbers), such as 6.9226543 and 4.0358165. A simple way to map such a pair to a real number is to interleave their digits. Thus, the two numbers above become the real number 0.6490232568514635. All the digits of the two coordinates are used, each is used once, none is duplicated, and none is deleted. It is clear that a different point would be mapped to a different real number and that every point can be mapped in such a way, which shows that our mapping is one-to-one.

We are familiar with the concepts of successor and predecessor. An integer N has both a successor $N + 1$ and a predecessor $N - 1$. Cantor showed that the rational numbers are countable; each can be assigned an integer subscript or tag. Thus, each rational number can be said to have a successor and a predecessor. The real numbers, again, are different. Given a real number a, we cannot point to its successor. If we find another real number b that may be the successor of a, then there is always another number, namely $(a + b)/2$, that is located between a and b and is therefore closer to a than b is. Thus, we say that a real number DOES NOT have an immediate successor or a predecessor; it does not have any immediate neighbors. The reals are so dense everywhere that they form a continuum, even though each real number corresponds to a point! We cannot imagine any collection of points, numbers, or any objects that are everywhere (extremely) dense but do not feature a predecessor/successor relation. The real numbers are therefore very counterintuitive.

The concept of a continuum is baffling, so here is an attempt to illuminate it. Imagine the interval $[0, 1]$. This is a closed interval that contains all the real numbers between 0 and 1, including the endpoints 0 and 1. Now remove the two endpoints to obtain the open interval $(0, 1)$. Starting at the midpoint 0.5, move steadily to the right, passing larger and larger numbers. We meet numbers such as 0.6, 0.8, 0.9, 0.95, 0.99, 0.999, and so on, but regardless of how much time we spend going through those numbers and regardless of how many numbers we pass, we will never reach the right end of the interval; we can never get to the last, largest number at the right end of the interval, because there isn't such a number! The same is true of the left end of the interval. If we move toward it, we reach smaller and smaller numbers, we move steadily toward zero, but we can never reach the left end of the interval because that end (the smallest positive real number) does not exist.

Here is another unexpected result. Pick up two real numbers x and y at random (but with a uniform distribution) in the open interval $(0, 1)$, divide them to obtain the real number $R = x/y$, and examine the integer I nearest R. We intuitively feel that I can be even or odd with the same probability, but careful calculations [Weisstein-picking 07] show that the probability of I being even is $0.46460\ldots$ instead of the expected 0.5. See Page 396 for the complete explanation.

This book contains text, tables, mathematical expressions, and figures, and it can be stored in the computer as a PDF file. Such a file, like any data file, can be considered an integer or a long string B of digits (decimal, binary, or to any base). A real number is also a (finite or infinite) string of digits. Thus, it is natural to ask, is there a real

number that includes B in its string of digits? The answer is yes. If $B = b_0b_1\ldots b_n$, then $0.b_0b_1\ldots b_n$ is such a number, and there are more such numbers.

Even more, there is a real number that includes in its infinite expansion all the books ever written and all those that will ever be written. Simply generate all the integers (we use binary notation) 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, 110, 111, 0000, 0001, ... and concatenate them to construct a real number R. From its construction, R includes every possible bitstring and thus every past and future book. (Students pay attention. Both the questions and answers of your next examination are also included in this number. It's just a question of finding this important part of R.) This R does not even have to be large. We can write it as $R = 0.01|00011011|000\ldots111|0000\ldots$, which makes it less that 1.

The term "lexicon" generally refers to a dictionary, but in mathematics, a Lexicon is a real number that contains in its expansion infinitely many times anything imaginable and unimaginable, everything ever written, or that will ever be written, and any descriptions of every object, process, and phenomenon, real or imaginary. Contrary to any intuitive feelings that we may have, such monsters are not rare. The surprising result, due to [Calude and Zamfirescu 98], is that almost every real number is a Lexicon! This may be easier to comprehend by means of a thought experiment. If we put all the reals in a bag, and pick out one at random, it will almost certainly be a Lexicon.

Champernowne constant is an example of a lexicon. The digits of this real number are the concatenation of all the natural numbers. Thus

$$C_{10} = 0.12345678910111213141516\ldots_{10} = 0.11011100101110111\ldots_2\,.$$

This number, which is also transcendental, was proposed by David Champernowne in 1933.

It is easy to come up with similar lexicons, such as a permutation of all the natural numbers, the digits of the squares of all the natural numbers, and the digits of all the prime numbers.

Gregory Chaitin, the originator of algorithmic information theory, describes in *The Limits of Reason* [Chaitin 07], a real number, denoted by Ω, that is well defined and is a specific number, but is impossible to compute in its entirety. This number is now known as Chaitin's constant.

Here is another baffling property of real numbers, known as the repeating 9's. The repeating real number $0.999\ldots$—which is also denoted by $0.\bar{9}$, $0.\dot{9}$, and $0.(9)$—is especially interesting. At first look it seems to be less than 1. However, it is easy to show (and to prove rigorously in many ways) that this real number equals 1. Here are several simple proofs:

1. Denote $a = 0.\bar{9}$. The number a cannot be greater than 1, so it must be either less than 1 or equal to 1. Assume that $a < 1$, then there must be numbers between a and 1, such as, for example, $b = (a + 1)/2$. A little thinking shows that b cannot exist, because there is no way to write it. The number a already uses the digit 9, which is the largest decimal digit, so b cannot use any larger digits. Also, a uses infinitely many 9's, so b cannot have any more 9's in its representation. Thus, b cannot exist, which implies that a cannot be less than 1 and must therefore be 1.

2. Using fractions, it is easy to see that

$$1 = \frac{3 \times 1}{3} = 3 \times \frac{1}{3} = 3 \times 0.\bar{3} = 0.\bar{9}.$$

3. Manipulating digits is also a simple procedure that proves our claim. Denote $a = 0.\bar{9}$, we obtain $10a = 9.999\ldots$, and $10a - a = 9.999\ldots - 0.999\ldots$ or $9a = 9 \Rightarrow a = 1$.

4. A geometric series has the form $ar + ar^2 + ar^3 + \cdots$. It is known that if the absolute value of r is less than 1, the sum of the series is $ar/(1-r)$. Our number $0.\bar{9}$ can be written as the geometric series

$$9\left[\frac{1}{10}\right] + 9\left[\frac{1}{10}\right]^2 + 9\left[\frac{1}{10}\right]^3 + \cdots$$

where r is 0.1. Its sum is therefore the finite quantity

$$\frac{9\left[\frac{1}{10}\right]}{1 - \frac{1}{10}} = 1.$$

5. The number $0.\bar{9}$ can be written as the limit of the infinite sequence

$$S_1 = 0.9 = 1 - \left(\frac{1}{10}\right)^1,\ S_2 = 0.99 = 1 - \left(\frac{1}{10}\right)^2,\ S_3 = 0.999 = 1 - \left(\frac{1}{10}\right)^3, \ldots,$$

whose limit is 1.

6. The following wrong proof shows that $0.\bar{9}$ is not 1. We start with the sequence 0.9, 099, 0.999, ... that converges to $0.\bar{9}$. Table 2.25 lists powers of this sequence and it is clear that each row of this table converges to a power of $0.\bar{9}$, while each column converges to zero. We therefore conclude that $\lim_{n\to\infty} 0.\bar{9}^n$ is zero. However, all powers of 1 are ones, so if $0.\bar{9}$ equals 1, all its powers must also equal 1. Thus, $0.\bar{9}$ cannot be 1.

$.9$	$.99$	$.999$	$\cdots$	$0.\bar{9}$
$.9^2$	$.99^2$	$.999^2$	$\cdots$	$0.\bar{9}^2$
$.9^3$	$.99^3$	$.999^3$	$\cdots$	$0.\bar{9}^3$
			$\vdots$	
$.9^n$	$.99^n$	$.999^n$	$\cdots$	$0.\bar{9}^n$

Table 2.25: Powers of 0.999....

The error is the result of swapping two limits. We show that $\lim_{m\to\infty} 0.\underbrace{99\ldots9}_{m}{}^n = 0.\bar{9}^n$ and then determine $\lim_{n\to\infty} 0.\bar{9}^n$ by writing $\lim_{n\to\infty}\lim_{m\to\infty} 0.\underbrace{99\ldots9}_{m}{}^n$, and computing this double limit by swapping the limits

$$\lim_{m\to\infty}\left(\lim_{n\to\infty} 0.\underbrace{99\ldots9}_{m}{}^n\right) = \lim_{m\to\infty} 0 = 0,$$

which causes the error.

These proofs demolish a long-held intuitive belief among students of mathematics, namely that a real number has one representation. It is now clear that infinitely many numbers have multiple representations. Thus, 52.8357 can also be written as the infinite-repeating number $52.8356\bar{9}$ and 1/3 also equals $0.\bar{3}$, 2/6, 4/12, and so on.

Insert: How to convert repeating digits to fractions.

Even if you doubt the proofs that $0.9\overline{9}$ equals 1, you might be interested to know how to convert other infinitely repeating decimals to fractions. Here are several examples, and more are available at reference [Mr H Tutoring 25].

- Given the repeating fraction $0.6\overline{6}$, simply multiply it by 10 and subtract

$$\begin{aligned} x &= 0.666666\ldots \\ 10x &= 6.666666\ldots \\ 9x &= 6 \\ x &= \frac{6}{9}. \end{aligned}$$

- The case of repeating strings of digits is similar. Given $0.\overline{274}$, multiply it by 1000 and subtract

$$\begin{aligned} x &= 0.274274274\ldots \\ 1000x &= 274.274274\ldots \\ 999x &= 274 \\ x &= \frac{274}{999}. \end{aligned}$$

The shortcut in these cases is therefore: Given an infinitely repeating string of n digits, generate the fraction whose numerator is the string and whose denominator is $n-1$.

- In more complex cases, the principle is to first separate the repeating string from the rest of the expression. Thus, $3.\overline{14}$ should be separated to $3+0.\overline{14}$, which is easily converted to $3\frac{14}{99}$. An expression such as $4.7\overline{25}$ should first be rewritten as $4.7+0.0\overline{25}$ and then as $4.7+0.\overline{25}/10$, which is easily converted to $4.7+(25/99)/10$. Finally, $3.71\overline{850}$ should be changed to $3.71+(850/999)/100$.

The following joke (from [Renteln and Dundes 05]) has long become a permanent part of mathematical lore:

Q: How many mathematicians does it take to screw in a lightbulb?

A: 0.999999....

The following simple problem, from `t.ly/8vjkJ` (YouTube channel `Prime Newtons`), further illustrates the close relation between the integers and the reals. We want to find all the real numbers x for which the expression $x+1/x$ is an integer. The first step is to assume that the real expression $x+1/x$ equals an integer K, and to find out

if such integers exist and what values they can take. We therefore write $x + 1/x = K$ and rearrange it as the quadratic equation

$$x^2 - Kx + 1 = 0, \tag{2.11}$$

with x as the unknown. In oder for this equation to have a real solution x, its discriminant $\sqrt{(-K)^2 - 4 \cdot 1 \cdot 1}$ must be nonnegative, which yields the relation $K^2 - 4 \geq 0$ or $|K| \geq 2$.

Thus, for $x + 1/x$ to be an integer, K must be in one of the intervals $[-\infty, 2]$ and $[2, \infty]$. Assuming that K is in the proper interval, the solution of Equation (2.11) is

$$x = \frac{K \pm \sqrt{K^2 - 4}}{2}.$$

Testing this for several integers in the proper intervals can easily be done either by hand or with the Mathematica code

```
 K = 33;
x = (K + Sqrt[K^2 - 4])/2;
x + 1./x
```

It is also obvious that for K in other intervals, the solutions are complex numbers.

Finally, consider the unusual base-13 function f, described on Page 433, which has the following property: When applied to all the real numbers in any real interval (a, b), this function returns every real number.

This is referred to as a stronger intermediate property and can also be described by saying that given any real interval (a, b) and a point c in it, we can choose any real number r and then construct a base-13 number R such that $f(R) = r$.

Unusual, unexpected, counterintuitive. The real numbers are weird, which is why we may also consider them beautiful!

2.7 Complex Numbers

The fundamental theorem of algebra, proved by Gauss in 1799, states that a polynomial of degree n has exactly n roots. However, it is easy to see that the simple polynomial $x^2 + 1$ (a parabola), which is of degree 2, has no roots, because $x^2 = -1$ implies $x = \pm\sqrt{-1}$, and such a number does not exist. When a number, any number, positive or negative, is squared, the result is either zero or positive; it cannot be negative.

This argument is one way to introduce the complex numbers. We simply say that in addition to all the other numbers that we know and use, which from now on will be called real, there is another number, the $\sqrt{-1}$, that will be denoted by i and will be referred to as imaginary (a term coined by Descartes, although Gauss proposed the term "lateral," meaning relating to the sides of something). It is all a matter of definition, and new mathematical definitions, objects, and theories can be freely introduced and explored, especially since they may provide new insights into mathematics or other disciplines, and provided they are consistent and do not give rise to contradictions or paradoxes.

Even more. Since people in general, and mathematicians in particular, like to extend and generalize new ideas and concepts, it seems natural to extend the new imaginary number and have compound (or hybrid) numbers that consist of both real and imaginary parts, and that are known as complex. Reference [WikiComplex 23] is a general, detailed introduction to this topic, complete with many applications of complex numbers.

Every high-school student "knows," when learning about i, that of course it is impossible to compute the square root of a negative number, and the imaginary i is only "supposed" to be such a square root. Later, however, perhaps at a university or on the job, the student learns to use i in calculations and designs that result in instruments, devices, and machines that really work! Gradually, i loses its mystery, it becomes a familiar tool or concept, and is routinely used every day.

Unfortunately, once the student gets used to working with i regularly, he often becomes oblivious to the obvious question of why does such an imaginary quantity work in our real world? This mystery is a special case of the general enigma of why mathematics, a creation of the human mind, is so effective in discovering and describing the world around us.

A complex number, usually denoted by z, is defined as a pair $z = a + bi$ where a and b are real numbers and $i = \sqrt{-1}$. Any product of the form $b\,i$ where b is real is referred to as an imaginary number.

History of Complex Numbers

Complex numbers first appeared when mathematicians became interested in solving equations. A degree-1 equation is of the form $ax + b = 0$, with a solution $x = -b/a$, but already a degree-2 (quadratic) equation, whose form is $ax^2 + bx + c = 0$, may exhibit strange behavior. The solution to such an equation is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

and this is undefined when the discriminant $b^2 - 4ac$ is negative. (Discriminant: A characteristic that enables things to be distinguished from one another.) The solutions to cubic (degree-3) equations suffer even more from the same problem.

In the 1530's, Italian mathematicians—among them Gerolamo Cardano, Niccolò Tartaglia, and Scipione del Ferro—came up with elaborate equations for solving certain types of cubic equations, and those equations made heavy use of square roots (and square roots of square roots) that often became negative. While initially refusing to admit the possibility of imaginary and complex numbers, those mathematicians slowly accepted them, wrote about them, and discovered their properties. It was their work that brought legitimacy to complex numbers and turned them into bona fide mathematical objects.

However, the person who is generally recognized as the "father" of imaginary and complex numbers is Rafael Bombelli. He was the first to believe in their usefulness and he published rules for operating on such numbers. If complex numbers are commonly used today in mathematics (both pure and applied), in science, and in engineering, it is to a large part due to his work.

One hundred years later, by the 1600's, complex numbers and especially the special number i, became part of mainstream mathematics. The little-known de Moivre formula $(\cos x + i \sin x)^n = \cos(nx) + i \sin(nx))$ was introduced by Abraham de Moivre in 1722. It was finally extended for any real n and also proved by Euler in 1749, using Euler's formula.

Today, it is common to visualize a complex number as a two-dimensional point in the so-called Argand diagram (Figure 2.26). This extremely useful graphics aid, the brainchild, in 1806, of Jean-Robert Argand, had first been published in 1799 by the land surveyor and mathematician Caspar Wessel in the little-known *Proceedings of the Copenhagen Academy*, where it languished unnoticed. It had also been anticipated by John Wallis in 1685.

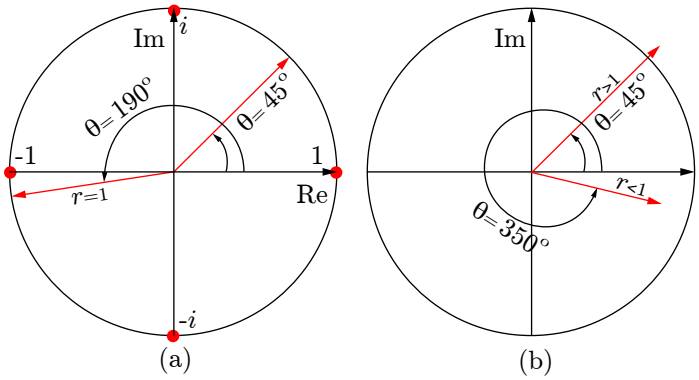


Figure 2.26: The Argand Diagram.

Finally, in the early 1800's, the work of Augustin-Louis Cauchy and Bernhard Riemann has made the branch of complex analysis a mature and well accepted part of modern mathematics.

Operations on Complex Numbers

The following arithmetic operations on complex numbers are immediate extensions of the same operations on real numbers $(a + bi) \pm (c + di) = \big(a \pm c, (b \pm d)i\big)$ and $(a + bi) \cdot (c + di) = ac + adi + bic + bidi = \big((ac - bd) + (ad + bc)i\big)$. Division of complex numbers $z1/z2$ is done by multiplying $z1$ by the conjugate of $z2$. A complex number $z = (a + bi)$ has no inverse, but has a conjugate $\bar{z} = a - bi$ (sometimes also denoted by z^*) which is the reflection of the number about the real axis (Re) of the Argand diagram.

Once this diagram and polar coordinates are introduced, it becomes clear that a complex number and its conjugate can also be written as $z = e^{ix}$ and $\bar{z} = e^{-ix}$, respectively. Also, the product $z \cdot \bar{z}$ is the real number $a^2 + b^2$ or r^2.

The last product is nicely illustrated by Figure 2.27. The conjugate $\bar{z} = e^{-ix}$ of the complex number $z = e^{ix}$ can be visualized as a rotation of z through an angle $2x$ about the real axis. Similarly, the product $z \cdot \bar{z}$ is the real number produced by adding the angles x and $-x$ to obtain zero, and multiplying the lengths l of the two numbers to obtain l^2.

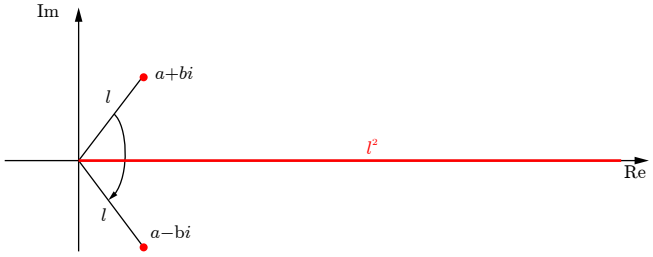


Figure 2.27: The Product $z \cdot \bar{z}$.

Useful relations and tools for dealing with complex numbers.

The relation $i^2 + 1 = 0$ implies the identities $i^{4k} = 1$, $i^{4k+1} = i$, $i^{4k+2} = -1$, and $i^{4k+3} = -i$, which hold for all integers k.

Euler's formula (or equation) $e^{ix} = \cos x + i \sin x$, introduced by Euler in 1748, is useful because it relates complex numbers to trigonometry, thereby simplifying operations on and processing of complex numbers. Euler proved it by formally manipulating complex power series, but the proof shown here is much simpler. It is based on the inverse relation $e^{\ln x} = x$ between e and the natural logarithm. We start with the as-yet-unknown complex number

$$z \stackrel{\text{def}}{=} \cos\theta + i \sin\theta$$

and prove that z must be our friend e. The first step is to differentiate z, which produces

$$\frac{dz}{d\theta} = -\sin\theta + i\cos\theta = i\cos\theta + i^2 \sin\theta = i(\cos\theta + i\sin\theta) = iz,$$

from which we derive $\frac{dz}{z} = id\theta$. When this is integrated, it yields

$$\int \frac{dz}{z} = \int id\theta,$$

or $\ln z = i\theta$, which we convert to $e^{\ln z} = e^{i\theta} + C$, and which in turn gives us $z = e^{i\theta} + C$.

The last step is to show that the constant C must be zero, which is done by setting $\theta = 0$. When θ is zero, both our original unknown number z and $e^{i\theta}$ become unities $\cos(0) + i\sin(0) = 1$ and $e^{i\cdot 0} = 1$, indicating that C is zero and $z = e^{i\theta} + C$ is therefore reduced to $z = e^{i\theta}$ to become our old trusted acquaintance.

The Argand diagram, published in 1806 by Jean-Robert Argand (although described by others before him) is an important graphical tool which illustrates the important relation between the Cartesian coordinates (a, b) of a complex number and its polar coordinates (r, θ), and implies the transformations $(a, b) = (r\cos\theta, r\sin\theta)$ and $(r, \theta) = \left(\sqrt{a^2 + b^2}, \arctan\frac{b}{a}\right)$. These transformations, combined with Euler's formula, also allow us to write

$$z = x + yi = |z|(\cos\theta + i\sin\theta) = |z|e^{i\theta},$$

where $|z| = \sqrt{x^2+y^2}$, $\theta = \arg(z) = \frac{1}{i}\ln\frac{z}{|z|}$, where $|z|$ is the absolute value of z and θ is the argument of z, which is normally used in the interval $[0, 2\pi]$.

• (A messy example). What is the general complex exponentiation $(a+bi)^{(c+di)}$? The derivation below assumes that the following relations are known: $r = e^{\ln r}$ and $r^{di} = e^{di\ln r}$ where r is one of the polar coordinates of $(a+bi)$. These relations justify the step marked by *.

$$\begin{aligned}(a+bi)^{(c+di)} &= \left[re^{i\theta}\right]^{c+di} = \left[re^{i\theta}\right]^{c}\left[re^{i\theta}\right]^{di} \stackrel{*}{=} r^c e^{ic\theta} r^{di} e^{-d\theta} = r^c e^{-d\theta} e^{i(c\theta + d\ln r)} \\ &= r^c e^{-d\theta}\left(\cos(c\theta + d\ln r) + i\sin(\cos(c\theta + d\ln r)\right).\end{aligned}$$

A correct, but by no means beautiful expression.

• At this point it is natural to ask what is the square root of an arbitrary complex number $a+bi$? This derivation is presented here, even though it is messy, because it will be needed on Page 145 for the exercise on infinity. We assume that this square root would itself be complex and we use the notation $\sqrt{a+bi} = x+yi$. When both sides of this are squared $\sqrt{a+bi}^2 = (x+yi)^2$, it is obvious that $a = x^2-y^2$ and $bi = 2xyi$, which implies $2xy = b \Rightarrow y = \frac{b}{2x}$ and $x^2 - y^2 = a \Rightarrow x^2 - \left[\frac{b}{2x}\right]^2 = a$. Multiplying both sides by $4x^2$ produces the quadratic equation $4(x^2)^2 - 4ax^2 - b^2 = 0$, whose messy solution is

$$x^2 = \frac{4a + \sqrt{16a^2+16b^2}}{8} \Rightarrow x = \pm\sqrt{\frac{a+\sqrt{a^2+b^2}}{2}}.$$

There is no minus sign in $4a + \sqrt{\cdots}$, because the value of x^2 must be nonnegative.

◇ **Exercise 2.27:** Derive the expression for y in a similar way.

The final result is

$$\sqrt{a+bi} = \pm\sqrt{\frac{a+\sqrt{a^2+b^2}}{2}} \pm i\sqrt{\frac{-a+\sqrt{a^2+b^2}}{2}}. \tag{2.12}$$

Notice that the relation $2xy = b$ implies $b > 0 \Rightarrow x\cdot y > 0$ and $b < 0 \Rightarrow x\cdot y < 0$. The case $b = 0$ means that $a+bi$ is real.

The nth Roots of Unity

A typical dictionary definition of "unity" is "the quality or state of being made one (unification)." In mathematics, the nth root of unity is any number z such that $z^n = 1$ for a positive integer n. However, it is easier to understand this concept by considering the equivalent definition which says: the nth roots of unity are the roots of the polynomials of the form $x^n - 1$. For $n = 2$, we factor the equation $x^2 - 1 = 0$ into $(x+1)(x-1) = 0$ to immediately obtain the two roots $+1$ and -1. However, for larger n, things quickly get more complicated. For $n = 4$, the roots of unity are the solutions to $x^4 - 1 = 0$, so we factor $(x+1)(x-1)(x^2+1)$ to immediately obtain the three roots 1, -1, and i. A little thinking may convince the reader that $-i$ is also a solution because $(-i)^2 = i^2 = -1$.

It turns out that most roots of unity are complex numbers, and are governed by the fundamental theorem of algebra. This important statement proves that every nth-degree polynomial has exactly n complex roots (although some of them may have a zero imaginary part and may therefore be real). As a result, we conclude that the equation $x^n = 1$ has n complex solutions, which are nth roots of unity and also constitute ALL the nth roots of unity. Figure 2.28 shows examples of the roots for several values of n.

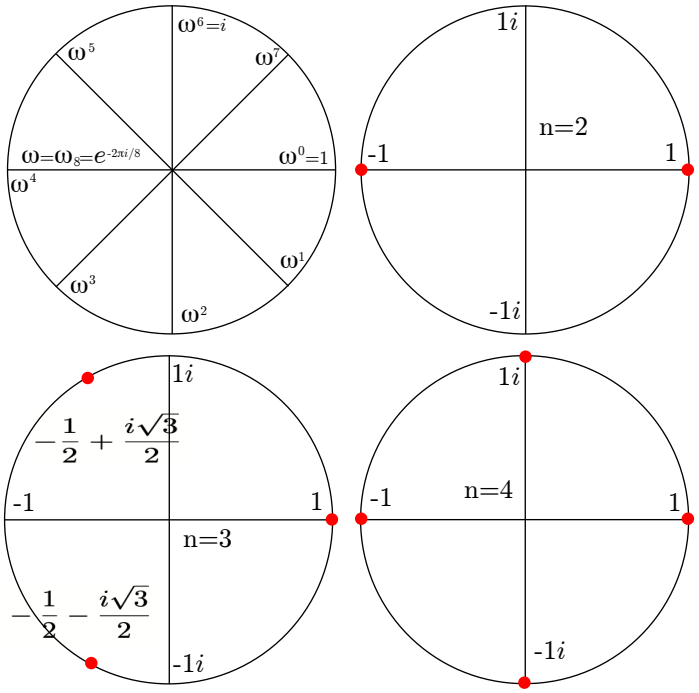


Figure 2.28: Several Roots of Unity.

The figure suggests that the roots are uniformly distributed over the unit circle (even if they don't form opposite pairs) and that the complex number $1+0i$ is always a root. The angle between consecutive roots on the circle is $2\pi/n$, so we define $\omega = e^{2\pi i/n}$ and spread the roots over the circle at points that correspond to powers of ω. Another interesting algebraic property of the roots is that their sum is zero for any n.

◇ **Exercise 2.28:** What are the roots of unity for $n = 8$?

◇ **Exercise 2.29:** Plot several roots of a general complex number (r, θ).

A note on how to multiply square roots of negative numbers. We know that in a mathematical expression, with numbers, letters, and arithmetic operations, the rules of operator precedence are: functions (including square roots), exponentiation, multiplication/division, and addition/subtraction. But what about square roots of negative numbers? With positive numbers, we have the simple identity $\sqrt{a \cdot b} = \sqrt{a}\sqrt{b}$, and this

can be applied to negative numbers by $\sqrt{-10} = \sqrt{(-1)\cdot 10} = \sqrt{-1}\sqrt{10} = i\sqrt{10}$. Thus, for example, $\sqrt{-16}\cdot\sqrt{-2} = i\cdot\sqrt{16}\times i\cdot\sqrt{2} = -\sqrt{32}$.

- Fractional powers. We now turn to fractional powers m/n of complex numbers. The starting point is $i = i^1 = i^{\frac{4}{4}} = (i^4)^{\frac{1}{4}} = 1^{\frac{1}{4}} = 1$, which is clearly wrong. The key to understanding the problem is how to interpret it. Instead of considering it as "i equals *the* fourth root of 1," we should think of it as "i is one of the fourth roots of unity," (the four fourth roots are 1, -1, i, and $-i$).

In an attempt to understand why we got just one result, we try to compute $i^{\frac{4}{4}}$ in two similar ways by swapping the exponents. We calculate it first as $(i^4)^{\frac{1}{4}}$ and then as $(i^{\frac{1}{4}})^4$. We use the relation $i^4 = 1$ and also notice that the polar coordinates of 1 as a complex number are $(r,\theta) = (1,0)$, but also $(1,2\pi)$, $(1,4\pi)$, and in general $(1,2n\pi)$ for $n = 0,1,2,\ldots$. Here is the first case

$$(i^4)^{\frac{1}{4}} = (1)^{\frac{1}{4}} = (e^{i2n\pi})^{\frac{1}{4}} = e^{i\frac{n}{2}\pi}.$$

For $n = 0,1,2$, and 3 this yields the correct, expected result namely, the four fourth roots of unity $e^0 = 1$, $e^{i\pi/2} = i$, $e^{i\pi} = -1$, and $e^{i3\pi/2} = -i$.

Next, we apply the same technique, using polar coordinates, to compute $(i^{\frac{1}{4}})^4$. We notice that the polar coordinates of i are $(1,2\pi)$ or, in general $(1,2\pi+2n\pi)$, which allows us to write

$$(i^{\frac{1}{4}})^4 = \left[\left(e^{i\frac{4n+1}{2}\pi}\right)^{\frac{1}{4}}\right]^4 = \left[e^{i\frac{4n+1}{8}\pi}\right]^4.$$

For $n = 0,1,2,3$ this yields the four quantities $(e^{i\pi/8})^4 = e^{i\pi/2} = i$, $(e^{i5\pi/8})^4 = e^{i5\pi/2} = i$, $(e^{i9\pi/8})^4 = e^{i9\pi/2} = i$, and $(e^{i13\pi/8})^4 = e^{i13\pi/2} = i$.

The obvious conclusion is that swapping the order of the exponents has produced the single, wrong result i. For a better understanding we should wait until Exercise 2.30 is worked out.

◇ **Exercise 2.30:** Compute the values of $i^{3/4}$ both as $(i^3)^{\frac{1}{4}}$ and as $(i^{\frac{1}{4}})^3$.

After reading the examples above and solving Exercise 2.30, the following conclusion should come as no surprise. When trying to raise an arbitrary complex number Z to the fractional power m/n, we must distinguish two cases. If the GCD (greatest common divisor) of m and n is 1, the order of the exponents does not matter. If, however, m and n have a common divisor, they should either be reduced first or the fractional power should be computed in the form $(Z^{\frac{1}{n}})^m$ (first the denominator and then the numerator).

The fifth roots of unity and the golden ratio.

To compute the fifth roots of unity, we need to solve the complex equation $z^5 = 1$. We start with the de Moivre formula

$$z = r(\cos\theta + i\sin\theta), \tag{2.13}$$

which was introduced by de Moivre in 1722 and was later extended by Euler for any real n, using his well-known formula. Armed with these results, we can rewrite Equa-

tion (2.13)

$$z^{1/n} = r^{1/n}\left\{\cos\left[\frac{\theta+2\pi k}{n}\right] + i\sin\left[\frac{\theta+2\pi k}{n}\right]\right\}, \quad \text{for } k = 0, 1, 2, \ldots, n-1.$$

Substituting $r = 1$ and $\theta = 0$, this reduces to

$$z^{1/n} = \cos\left[\frac{2\pi k}{n}\right] + i\sin\left[\frac{2\pi k}{n}\right] = e^{\frac{2\pi k}{n}i}. \tag{2.14}$$

Applying Equation (2.14) to $(1/n) = 5$, we obtain the five roots, for $k = 0, 1, 2, 3,$ and 4

$$\begin{aligned}
k = 0\colon z^{1/5} &= \cos\frac{0\pi}{5} + i\sin\frac{0\pi}{5} = e^{\frac{0\pi}{5}i} = 1,\\
k = 1\colon z^{1/5} &= \cos\frac{2\pi}{5} + i\sin\frac{2\pi}{5} = e^{\frac{2\pi}{5}i},\\
k = 2\colon z^{1/5} &= \cos\frac{4\pi}{5} + i\sin\frac{4\pi}{5} = e^{\frac{4\pi}{5}i},\\
k = 3\colon z^{1/5} &= \cos\frac{-4\pi}{5} + i\sin\frac{-4\pi}{5} = e^{\frac{-4\pi}{5}i},\\
k = 4\colon z^{1/5} &= \cos\frac{-2\pi}{5} + i\sin\frac{-2\pi}{5} = e^{\frac{-2\pi}{5}i}.
\end{aligned}$$

In order to contain the five roots in the Argand diagram, we have to restrict θ to the interval $[-\pi, \pi]$. Bearing in mind that the cosine and sine have period 2π, we use for $k = 3$ and $k = 4$ the relations

$$\frac{6\pi}{5} - \frac{10\pi}{5} = \frac{-4\pi}{5} \quad \text{and} \quad \frac{8\pi}{5} - \frac{10\pi}{5} = \frac{-2\pi}{5}, \quad \text{respectively.}$$

The five roots are illustrated in the Argand diagram of Figure 2.29.

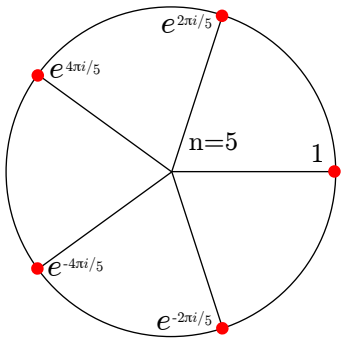


Figure 2.29: The Fifth Roots of Unity.

There are roots and then there are roots

It is important to point out that when it comes to roots, we sometimes have to distinguish the real and complex cases, and a good example of this unexpected behavior is the roots of negative numbers. If x is positive and n is a positive integer, then $\sqrt[n]{x}$ and $x^{1/n}$ are the same, but if x is negative, we have to consider complex roots.

From the previous discussion it is clear that the nth roots of a complex number z are given by

$$z^{1/n} = \left[re^{i(\theta+2\pi k)}\right]^{1/n} = \sqrt[n]{r}\exp\left[i\Big(\frac{\theta}{n} + \frac{2k\pi}{n}\Big)\right],$$

For $k = 0, 1, 2, \ldots, n-1$. There are n such roots, and the one corresponding to $k = 0$ is normally considered the principal root.

Thus, for negative x, the notation $\sqrt[n]{x}$ means to take the real root, if it exists, otherwise, the principal complex root. The notation $x^{1/n}$ means to compute the set of n complex roots, while sometimes retaining only the principal root.

For illustration, we consider the cube root of -1. If we are only interested in the real case, we can simple write $\sqrt[3]{-1} = (-1)^{1/3} = -1$, but if we need the complex roots, then we first realize that $z = (1, \pi) = e^{i\pi+2\pi k}$ and then compute

$$z^{1/3} = \sqrt[3]{1}\exp\left[i\Big(\frac{\pi}{3} + \frac{2k\pi}{3}\Big)\right].$$

The three roots for $k = 0, 1, 2$ are, respectively

$$\begin{aligned} e^{i\pi/3} &= \frac{1}{2} + \frac{i\sqrt{3}}{2}, \text{ principal root,} \\ e^{i\pi} &= (-1, 0), \text{ real number,} \\ e^{i5\pi/3} &= \frac{1}{2} - \frac{i\sqrt{3}}{2}, \text{ for } k = 2. \end{aligned}$$

◇ **Exercise 2.31:** (a) What are the three cube roots of -1? (b) Compute the three solutions of the equation $x^3 = 8$.

◇ **Exercise 2.32:** Can the sum of two imaginary numbers be real?

2.8 Hypercomplex Numbers?

Once we come to terms with complex numbers, their "existence," and their operations, the natural question is can we have multipart- or hypercomplex numbers, numbers that consist of three, four, or even more parts? This section shows why three-part numbers—which might be referred to as 3D complex, triternions, triernions, or trinions—cannot be defined in a way that makes sense. It is easy to show that such numbers cannot be multiplied in a way that would make them natural extensions of the two-dimensional complex numbers.

A complex number has the form $a + bi$, and the Argand diagram, with its real and imaginary axes, allows us to consider such a number a point in two-dimensional space. Therefore, a trinion should have the form $a+bi+cj$, and should be visualized as a point in a three-dimensional Argand diagram with real, i, and j axes that are perpendicular to one another. The real coefficients a, b, and c of such a number are the lengths (or scale factors) of the unit base vectors along the real, imaginary, and jth coordinate axes of this diagram.

> My train of thoughts was of this kind. Since $\sqrt{-1}$ is in a certain well-known sense, a line perpendicular to the line 1, it seemed natural that there should be some other imaginary to express a line perpendicular to both the former; ... Calling the old root, as the Germans do, i, and the new one j, I inquired what laws ought to be assumed for multiplying together $a + ib + jc$ and $x + iy + jz$. ...
> —[https://t.ly/H0mDv43], Letter from Hamilton to John Graves, October 17, 1843.

It is important to understand that the quantities i and j are perpendicular unit vectors and are therefore independent. (They are the same length, but point in independent directions.) Not only are they different, but they cannot be related. It is impossible, for example, to express j in terms of i. Also, since both are unit vectors and $i^2 = -1$, it is reasonable to assume that j^2 also equals -1.

The problem with operations on trinions becomes apparent when we try to multiply them. The first step in computing the product $(a+bi+cj)\times(d+ei+fj)$ is to open the parentheses. This results in a nine-term expression that starts with ad and ends with $cjfj$, and includes terms of the form $\ldots ij$. We now have to decide on the value of the simple product $i \cdot j$. In order for this multiplication to make sense, its value must be a real number r, or an imaginary number ri, or of the form rj, or an arbitrary trinion such as $k + li + mj$. We now show that none of these forms can be a candidate for $i \cdot j$.

- Assume that $i \cdot j = r$ and multiply both sides by i on the left. This results in $ii \cdot j = ir$ or $-j = ir$, i.e., j is an imaginary number, it is related to i.
- Assume that $i \cdot j = ir$ and multiply both sides by i on the left. This results in $ii \cdot j = iir$ or $-j = -r$, i.e., j is real.
- Assume that $i \cdot j = rj$ and multiply both sides by j on the right. This results in $i \cdot j \cdot j = rj \cdot j$ or $i = r$, i.e., i is real.

We assume that the product of the unit vectors i and j may perhaps be noncommutative, but we want to maintain the property of associativity, because otherwise our 3D hypercomplex numbers wouldn't be true extensions of the complex numbers and would not satisfy many of the basic properties of the real and complex numbers.

- The last case to be checked is $i \cdot j = a+bi+cj$, where a, b, and c are real. Multiplying this by i on the left yields

$$i \cdot i \cdot j = ai + iib + cij = ai - b + c(a + bi + cj),$$

or $-j = (ca-b)+(a+bc)i+c^2 j$. Since i and j are independent, this implies $ca-b = 0$, $a + bc = 0$, and $c^2 = 1$. However, already the relation $c^2 = 1$ implies that c is the imaginary i, but c is real. Also, if we substitute i for c in $a + bc = 0$ it implies that a, which must be real, is imaginary.

The conclusion is that the simple product $i \cdot j$ cannot be a point in the three-dimensional Argand diagram, because it cannot lie on any of its axes and cannot be an arbitrary point inside the diagram. Hence, trinions cannot be defined, in the sense that it is impossible to define operations on them that would be compatible with the operations on real and complex numbers.

Complex numbers and points

Complex numbers are similar to the two-dimensional points of Section 8.9, but there are the following important differences.

- The two components of a point (x, y) have their individual meanings. They describe the distances of the point from the two coordinate axes. Similarly, the two components of a complex number (a, b) are its distances from the real and imaginary axes, but they constitute one unit and must be treated differently when the number is operated on. We must keep in mind that the second component b is a multiple of the imaginary $i = \sqrt{-1}$, because this affects the interpretation of the results of our computations.

- Complex numbers are numbers and can be operated on. Specifically, they can be added and subtracted. Points can be subtracted, but adding points is meaningless, because the sum depends on the particular choice of coordinate system, as discussed in Section 8.9.

- Points exist in any number of dimensions, and can be treated and operated on regardless of how many components they have. Complex numbers, on the other hand, are two-dimensional and cannot easily be extended to more dimensions. It has taken William Rowan Hamilton years to realize that there is no reasonable way to define the multiplication of triplets, which is why he started thinking of complex 4-tuples, new objects that he named quaternions. He was finally able to define a reasonable, but not perfect, multiplication rule for those objects. His rule allowed for a reciprocal of a quaternion, but was noncommutative. Thus, something was lost when extending the two-dimensional complex numbers to the four-dimensional quaternions.

It has later been proved that it is possible to define reasonable rules of multiplication for hypercomplex numbers in spaces whose number of dimensions is a multiple of 2, but as we climb up to those higher dimensions, the multiplication rules lose more of the nice properties of real number multiplication. The multiplication of 8-dimensional complex numbers, octonions, is nonassociative, while that of sedenions, the 16-dimensional complex numbers, is noncommutative, nonassociative, and also lacks a property called alternativity. Sedenions also include zero divisors, which means that there exist non-zero sedenions a and b such that $ab = 0$.

◇ **Exercise 2.33:** What would be the name of 32-dimensional complex objects?

◇ **Exercise 2.34:** Why the term "complex numbers" and not duonions? Recall that complex numbers are doubles, so they can be compared with quaternions, which are quadruples, and octonions, which are octuples.

Along the same lines, why is the term "imaginary" used for $i = \sqrt{-1}$? After all, this quantity has many useful (real?) applications, and the term "imaginary" may mislead many to consider i a useless quantity.

Fun with complex numbers

The fun described in the next few paragraphs is based on the ideas in the video at [Mathoma 16], and on the following facts:

- The product of two complex numbers is generally a complex number.
- A complex number a can be represented as a point in the Argand diagram, which is why a product of the form $a \cdot b$ can be viewed as such a point.
- The absolute value $|a|$ of a complex number a is its distance from the origin of the Argand diagram.
- The absolute value $|a \cdot b|$ of a product equals the product $|a| \cdot |b|$ of the absolute values of a and b.

Now consider a complex number b whose absolute value is 1. In the Argand diagram, this number is located on the unit circle centered on the origin of the diagram. We now choose an arbitrary complex number a and observe that its distance from the origin is $|a|$, so it is located on the radius-$|a|$ circle centered on the origin. We multiply $c = a \cdot b$ and notice that $|c| = |a|$. Point c is therefore on the same circle as a, which implies that the multiplication $a \cdot b$ has moved a along its circle; it has rotated a about the origin.

Part (a) of Figure 2.30 shows 150 random points on the Argand diagram, and part (b) shows the effect of rotating each five times through four degrees. It is obvious that each point has become the start of a circular arc. The complex number $b = \cos(4Degree) + i\sin(4Degree)$ has absolute value 1, so the product $a \cdot b$ moves a four degrees along its circle.

Part (c) of the figure illustrates spiraling. A spiral can be motion in a circle whose radius grows steadily, and this effect is achieved by preparing a complex number such as $b = 1.02[\cos(4\text{ Degree}) + \mathrm{i}\sin(4\text{ Degree})]$ and multiplying a by powers of b. This is shown in the Mathematica code listings that accompany Figure 2.30.

Finally, part (d) of the figure shows how to scale a point a by multiplying it by successive powers of a real number b, such as 1.04.

Mathematics can be fun.

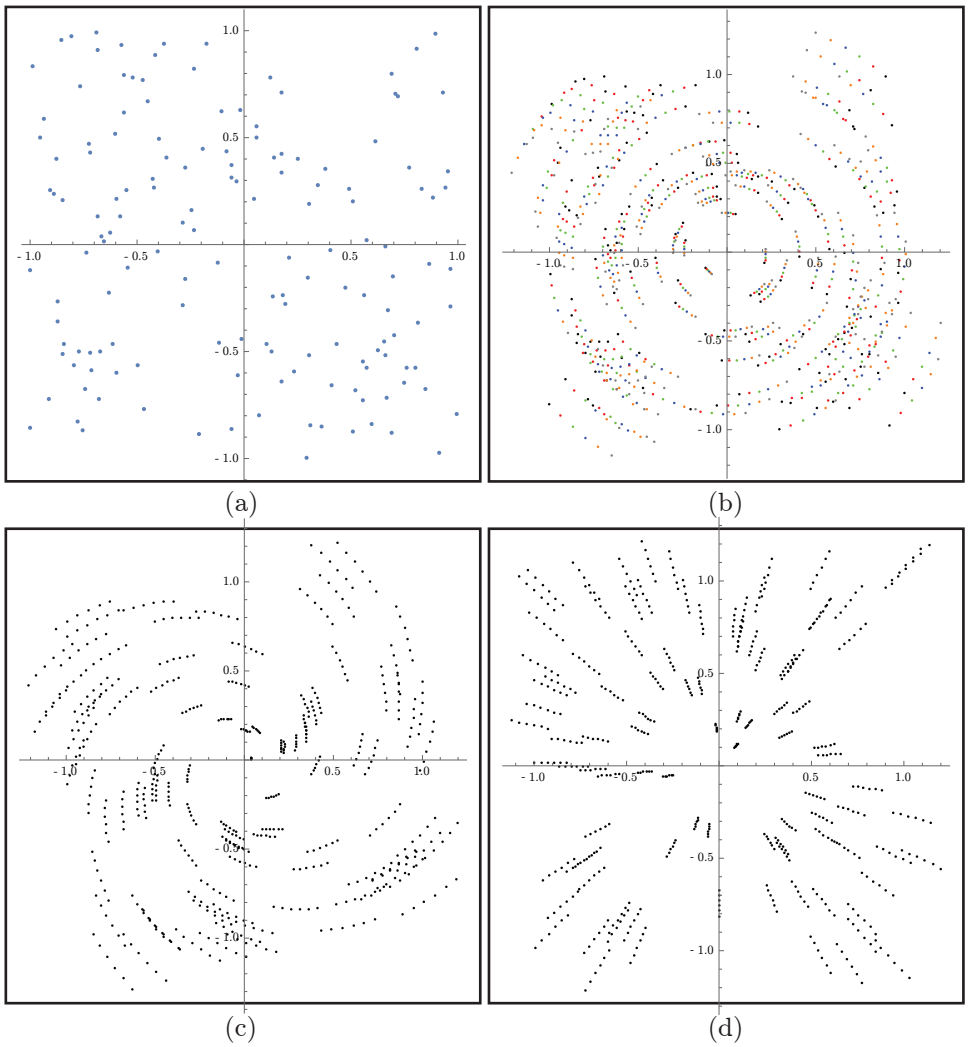


Figure 2.30: Fun with Complex Numbers.

2.9 Quaternions

Complex numbers can be considered points in the xy plane. The pair of real numbers that constitute the complex number (a, b) can be interpreted as the xy coordinates of a two-dimensional point (a, b). Section 2.8 shows that there is no reasonable way to define hypercomplex numbers (trinions) of the form (a, b, c) that could be interpreted as three-dimensional points, and this problem bothered the eminent Irish mathematician William Rowan Hamilton for years. He knew that multiplying complex numbers amounts to a rotation in two dimensions, so it made sense for him to ask whether multiplying the new hypercomplex numbers could be interpreted as a rotation in three dimensions.

```
l = {}; (*Rotating random complex points*)
b4 = Cos[4 Degree] + I Sin[4 Degree];
b8 = Cos[8 Degree] + I Sin[8 Degree];
b12 = Cos[12 Degree] + I Sin[12 Degree];
b16 = Cos[16 Degree] + I Sin[16 Degree];
b20 = Cos[20 Degree] + I Sin[20 Degree];
Do[AppendTo[l, RandomComplex[{-1 - I, 1 + I}]], {i, 150}]
ComplexListPlot[l]
ComplexListPlot[{l, l b4, l b8, l b12, l b16, l b20},
 PlotStyle -> {Black, Red, Green, Blue, Orange, Gray}]

l = {}; (*Spiraling random complex points*)
b = 1.02 (Cos[3 Degree] + I Sin[3 Degree]);
Do[AppendTo[l, RandomComplex[{-1 - I, 1 + I}]], {i, 100}]
ComplexListPlot[{l, l b3, l b^2, l b^3, l b^4, l b^5},
 PlotStyle -> {Black}]

l={}; (*Scaling random complex points*)
b=1.04;
Do[AppendTo[l,RandomComplex[{-1-I,1+I}]],{i,100}]
ComplexListPlot[{l,l b,l b^2,l b^3,l b^4,l b^5},
 PlotStyle->{{Black,AbsolutePointSize[2]}}]
```

Mathematica Listings for Figure 2.30

> Every morning in the early part of the above-cited month [Oct. 1843] on my coming down to breakfast, your brother William Edwin and yourself used to ask me, "Well, Papa, can you multiply triplets?" Whereto I was always obliged to reply, with a sad shake of the head, "No, I can only add and subtract them."
>
> —William Rowan Hamilton.

Hamilton could not come up with a reasonable rule for multiplying hypercomplex numbers that are triplets, and he eventually discovered, in October 1843, that he needed to add a fourth component to his triplets (i.e., turn them into 4-tuples) in order to multiply them in a way that made sense and obey (most of) the rules of conventional algebra. We don't know why he named these new entities quaternions and not quadruples or quadruplets. (The ancient meaning of quaternion is a group of four soldiers acting as a guard, specifically a Roman military unit.)

In modern notation, a quaternion $\mathbf{q}$ can be represented as a 2×2 matrix of complex numbers

$$\mathbf{q} = \begin{pmatrix} z & w \\ -w^* & z^* \end{pmatrix} = \begin{pmatrix} a+ib & c+id \\ -c+id & a-ib \end{pmatrix},$$

where z and w are complex numbers and a, b, c, and d are real. This can also be written (by analogy with the complex numbers $a\times 1 + b\times i$) as $\mathbf{q} = a\mathbf{U} + b\mathbf{I} + c\mathbf{J} + d\mathbf{K}$, where

$$\mathbf{U} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{I} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \mathbf{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{and } \mathbf{K} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

(Note that $\mathbf{U}$, not $\mathbf{I}$, is used here to denote the identity matrix. These matrices are closely related to the Pauli spin matrices used in particle physics.) From the definitions above, it follows that $\mathbf{I}^2 = -\mathbf{U}$, $\mathbf{J}^2 = -\mathbf{U}$, and $\mathbf{K}^2 = -\mathbf{U}$. We therefore conclude that $\mathbf{I}$, $\mathbf{J}$, and $\mathbf{K}$ are three different solutions of the matrix equation $\mathbf{X}^2 = -\mathbf{U}$ and should be considered the square roots of the negative of the identity matrix.

Quaternions can also be viewed as elements of a four-dimensional vector space, one of whose bases is given by

$$\mathbf{i} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \quad \mathbf{j} = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{k} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad \mathbf{1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

> And here there dawned on me the notion that we must admit, in some sense, a fourth dimension of space for the purpose of calculating with triplets; or transferring the paradox to algebra, must admit a third distinct imaginary symbol k, not to be confounded with either i or j, but equal to the product of the first as multiplier, and the second as multiplicand; and therefore was led to introduce quaternions, such as $a + ib + jc + kd$, or (a, b, c, d).
> —[`https://t.ly/H0mDv43`], Letter from Hamilton to John Graves, October 17, 1843.

Reference [Arianrhod 24] has a lively survey of the history of quaternions and much information about Hamilton and his life. For more on the history of quaternions and Hamilton's epiphany, see the letter from Hamilton to John Graves, dated October 17, 1843, available at `https://t.ly/H0mDv`, as well as the lively videos at `https://t.ly/7WXeB` and `https://t.ly/b0KEY`

After years of fruitless search for reasonably-behaved triplets of the form $a+ib+jc$, Hamilton finally hit on the right idea. As he later explained to his friend John Graves, "there dawned on me the notion that we must admit, in some sense, a fourth dimension of space." He therefore turned his attention to a new, four-dimensional complex number of the form $a + ib + jc + kd$, where i, j, and k were new, imaginary entities.

The next few paragraphs discuss the two new imaginary numbers j and k, introduced by Hamilton, in an attempt to familiarize the reader with these quantities, better understand their nature, and give them legitimacy. After dealing with complex numbers for a while, we feel that i is our friend because it relates to the familiar real numbers. It is also easy to accept the new imaginary j, because we can imagine it as a new z coordinate, perpendicular to the complex plane. Any point in the three-dimensional space defined by the real axis, imaginary axis, and the new z axis can be represented by the triplet $a + ib + jc$.

But what meaning can be assigned to the new imaginary k? Living in a three-dimensional world, we cannot visualize a fourth coordinate axis that would be perpendicular to the other three axes, but we can introduce both k and quaternions through

rotations in three dimensions. In fact, rotations in three dimensions are an important application of quaternions; so much so that such rotations are sometimes considered the definition of quaternions.

> Time is said to have only one dimension, and space to have three dimensions. ... The mathematical quaternion partakes of both these elements; in technical language it may be said to be "time plus space." or "space plus time:" And in this sense it has, or at least involves a reference to, four dimensions. ... And how the One of Time, of Space the Three, Might in the Chain of Symbols girdled be.
>
> —William Rowan Hamilton (c. 1853.

When discussing the mathematics of rotation, we start by considering the rotation of a point. In practice, however, we often want to rotate an entire object, and this is done by rotating several key points, and then reconstructing the rotated object from them.

Rotations in two dimensions are simple, because they are always done in the xy plane. They are also easy to visualize, because the center of rotation is usually the origin. These rotations are discussed on Pages 217 through 221. In contrast, rotations in three dimensions are much more difficult to visualize. Relatively few people posses the kind of geometric intuition that allows one to visualize three-dimensional objects and their transformations (rotations, reflections, and shearing).

> [William] Thurston, they say, is possessed of a geometric intuition unlike that of any other human. "When you see him or talk to him, he is often staring out into space and you can see that he sees these pictures," said John Morgan, a friend of Thurston's
>
> —Masha Gessen, *Perfect Rigour*, Icon Books, (2011).

The special case of three-dimensional rotations about the coordinate axes is easy to picture. Such a rotation can be fully specified by (1) the particular coordinate axis, (2) the angle of rotation, and (3) the direction of rotation, clockwise or counterclockwise, as seen when looking through the positive rotation axis to the origin. (The direction of rotation can also be implicitly specified by the sign of the rotation angle.) Specifying an arbitrary three-dimensional rotation is similar, with the difference that such a rotation is about a general axis in space. An axis is similar to a vector in the sense that it has both direction and magnitude, but unlike a vector, an axis also has a well-defined location in space. The difference between an axis and a vector is also discussed on Page 457.

Because of the difficulty of specifying an axis, a rotation in three dimensions is usually specified about a vector, as described on Page 457. This has the advantage that only four numbers are needed, one for the angle and three for the direction vector. Thus, a general rotation algorithm starts with a direction vector $\mathbf{k}$ (a unit vector) and a vector $\mathbf{v}$ to be rotated about $\mathbf{k}$. The result is a rotated vector $\mathbf{v}*$. If we denote the endpoints of $\mathbf{v}$ and $\mathbf{v}*$ by $\mathbf{P}$ and $\mathbf{P}^*$, respectively, then we can interpret these points as the start and end of the rotation, and we can translate them rigidly to any location in the coordinate space.

Another approach to three-dimensional rotations is to break such a rotation into three rotations about the coordinate axes. Once the three rotation angles have been determined, the complete rotation is performed in one of three ways as follows:

- Use the three rotation matrices of Equations (8.13) or (8.14) to rotate the point step by step about the coordinate axes.

- Multiply these rotation matrices to obtain the general three-dimensional rotation matrix, and use it to rotate the point in a single step.

- Convert the three rotation angles to a quaternion and use it to perform the rotation.

However, it is often unclear how to represent an arbitrary rotation into three angles. Also, in animation, it is crucial to interpolate smoothly between frames. Given two sets of rotation angles, it is not clear how to compute several intermediate sets that would represent frames corresponding to smooth rotations. Quaternions, however, provide an easy way to interpolate two three-dimensional directions. Just calculate the initial and final quaternions, and interpolate them linearly by means of Equation (8.20).

Adding/subtracting 4-tuples by simply adding/subtracting corresponding components is simple and doesn't raise any problems. What is not obvious is how to define quaternion multiplication such that the product of two quaternions would represent a combined rotation.

> Quaternions came from Hamilton after his really good work had been done; and, though beautifully ingenious, have been an unmixed evil to those who have touched them in any way, including Clerk Maxwell.
>
> —William Thomson, Lord Kelvin (1892).
>
> The quaternion was not only not required, but was a positive evil of no inconsiderable magnitude.
>
> —Oliver Heaviside.

We start by simply multiplying two 4-part quaternions to obtain the string of 16 partial products

$$\begin{aligned}
&(a_1 + b_1 i + c_1 j + d_1 k)(a_2 + b_2 i + c_2 j + d_2 k) = \\
&+a_1 a_2 + a_1 b_2 i + a_1 c_2 j + a_1 d_2 k \\
&+b_1 a_2 i + b_1 b_2 i^2 + b_1 c_2 ij + b_1 d_2 ik \\
&+c_1 a_2 j + c_1 b_2 ji + c_1 c_2 j^2 + c_1 d_2 jk \\
&+d_1 a_2 k + d_1 b_2 ki + d_1 c_2 kj + d_1 d_2 k^2.
\end{aligned}$$

The problem that faced Hamilton at this point was how to assign meaning to the two new imaginaries j and k, such that terms of the form j^2, ij, jk, and similar others would disappear, and only terms with i, j, and k would remain.

Hamilton's epiphany consisted of the four relations $\mathbf{i^2 = j^2 = k^2 = ijk = -1}$, from which he deduced, as shown here, the values of the remaining triplets and pairs that he needed:

$$\begin{aligned}
&ijk = -1 \rightarrow iijki = i(-1)i = -ii \rightarrow -jki = 1 \rightarrow jki = -1 \\
&ijk = -1 \rightarrow kijkk = -1 \rightarrow k(-1)k = -kk \rightarrow -kij = 1 \rightarrow kij = -1.
\end{aligned}$$

$$ijk = -1 \to iijk = -i \to jk = i,$$
$$jki = -1 \to jjki = -j \to ki = j,$$
$$kij = -1 \to kkij = -k \to ij = k.$$

$$ijk = -1 \to ijkkj = -kj \to ij(-1)j = -kj \to kj = -i,$$
$$jki = -1 \to jkiik = -ik \to jk(-1)k = -ik \to ik = -j,$$
$$kij = -1 \to kijji = -ji \to ki(-1)i = -ji \to ji = -k.$$

Based on these values of the three imaginaries, Table 2.31 shows the values of the six pairs of imaginaries, and how each can be obtained as the difference (cross product − dot product) of **i**, **j**, and **k**, when these imaginaries are considered vectors.

a×**b**	**i**	**j**	**k**
i	0	**k**	−**j**
j	−**k**	0	**i**
k	**j**	−**i**	0

−

a·**b**	**i**	**j**	**k**
i	1	0	0
j	0	1	0
k	0	0	1

=

ab	**i**	**j**	**k**
i	−1	**k**	−**j**
j	−**k**	−1	**i**
k	**j**	−**i**	−1

Table 2.31: Multiplying **i**, **j**, **k** as Vectors.

The quaternion multiplication above can now be written in terms of the three imaginary quantities

$$\begin{aligned}&(a_1 + b_1 i + c_1 j + d_1 k)(a_2 + b_2 i + c_2 j + d_2 k) = \\ &a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2 \\ &+(a_1 b_2 + b_1 a_2 + c_1 d_2 - d_1 c_2) i \\ &+(a_1 c_2 - b_1 d_2 + c_1 a_2 + d_1 b_2) j \\ &+(a_1 d_2 + b_1 c_2 - c_1 b_2 + d_1 a_2) k.\end{aligned}$$

The eight quaternions $\pm\mathbf{1}$, $\pm\mathbf{i}$, $\pm\mathbf{j}$, and $\pm\mathbf{k}$ form a group of order 8 with multiplication as the group operation.

It is common to interpret quaternions as a pair (scalar, vector). In this interpretation, a quaternion $\mathbf{q}$ can be represented as the sum $\mathbf{q} = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ or the pair $[s, \mathbf{v}]$, where $s = w$ and $\mathbf{v} = (x, y, z)$. This interpretation will be especially useful when we get to rotations, because we can then consider $\mathbf{v}$ the axis of rotation, a three-dimensional vector whose components are located along the axes defined by the basis vectors **i**, **j**, and **k**, and consider s the angle of rotation (or a quantity related to that angle). A quaternion where $s = 0$ is called a vector quaternion, because it is essentially just a vector.

The conjugate quaternion is given by $\mathbf{q}^* = w - x\mathbf{i} - y\mathbf{j} - z\mathbf{k}$. The sum or difference of two quaternions is the obvious

$$\mathbf{q}_1 \pm \mathbf{q}_2 = (w_1 + w_2) \pm (x_1 + x_2)\mathbf{i} \pm (y_1 + y_2)\mathbf{j} \pm (z_1 + z_2)\mathbf{k} = [s_1 \pm s_2, (\mathbf{v}_1 \pm \mathbf{v}_2)],$$

and the product is the nonobvious

$$\mathbf{q}_1 \cdot \mathbf{q}_2 = (w_1 w_2 - x_1 x_2 - y_1 y_2 - z_1 z_2) + (w_1 x_2 + x_1 w_2 + y_1 z_2 - z_1 y_2)\mathbf{i}$$

$$+ (w_1y_2 - x_1z_2 + y_1w_2 + z_1x_2)\mathbf{j} + (w_1z_2 + x_1y_2 - y_1x_2 + z_1w_2)\mathbf{k}$$
$$= [(s_1s_2 - \mathbf{v}_1 \bullet \mathbf{v}_2), (s_1\mathbf{v}_2 + s_2\mathbf{v}_1 + \mathbf{v}_1 \times \mathbf{v}_2)]. \quad (2.15)$$

Quaternion product is associative (i.e., $(\mathbf{q}_1\mathbf{q}_2)\mathbf{q}_3 = \mathbf{q}_1(\mathbf{q}_2\mathbf{q}_3)$) but not commutative. This makes sense because we are going to use quaternions to represent rotations in three dimensions, and such rotations are not commutative.

◇ **Exercise 2.35:** Show why rotations in three dimensions are not commutative.

An appropriate measure of the size of a quaternion is its norm, defined as

$$\| \mathbf{q} \| = \mathbf{q} \cdot \mathbf{q}^* = \mathbf{q}^* \cdot \mathbf{q} = \sqrt{w^2 + x^2 + y^2 + z^2} = \sqrt{s^2 + x^2 + y^2 + z^2}.$$

It is easy to verify that the norm is multiplicative, $\| \mathbf{q}_1\mathbf{q}_2 \| = \| \mathbf{q}_1 \| \| \mathbf{q}_2 \|$ (i.e., the norm of a product equals the product of the two individual norms). A unit quaternion, also called a versor is one for which $\| \mathbf{q} \| = 1$.

The inverse of a quaternion is given by

$$\mathbf{q}^{-1} = \frac{\mathbf{q}^*}{(\mathbf{q}\mathbf{q}^*)} = \frac{\mathbf{q}^*}{|\mathbf{q}|^2} = \frac{\mathbf{q}^*}{w^2 + x^2 + y^2 + z^2},$$

so quaternion division $\mathbf{q}_1/\mathbf{q}_2$ (except by zero) is performed by multiplying $\mathbf{q}_1$ by the inverse $\mathbf{q}_2^{-1}$. It's easy to verify that $\mathbf{q}\mathbf{q}^{-1} = [1, (0, 0, 0)] = [0, \mathbf{0}]$.

[Mathworld 25] and [WikiQuaternion 25] are basic references for quaternions. The interesting, unexpected connection between quaternions and rotations is developed in detail in [Hanson 06] (see especially page 50 of this reference).

◇ **Exercise 2.36:** (If four, why not more?) Quaternions are an extension of vectors. Are there extensions of quaternions?

> "Quaternion" was, I think, defined by an American schoolgirl to be "an ancient religious ceremony." This was, however, a complete mistake: The ancients—unlike Prof. Tait—knew not, and did not worship Quaternions.
>
> —Oliver Heaviside (1843).

Complex numbers and two-dimensional rotations

To illustrate the connection between quaternions and three-dimensional rotations, we first consider the application of complex numbers to two-dimensional rotations. Multiplying an arbitrary complex number $a + bi$ by i yields $(a + bi)i = ai - b$, and it is easy to see, by plotting both $a + bi$ and $-b + ai$ as points on the Argand diagram that they represent a 90° counterclockwise rotation about the origin. Also their dot product $(a, b) \cdot (-b, a) = -ab + ab$ is zero. A more general case of rotation is illustrated by Figure 2.32. Part (a) of the figure shows that rotating point $(1, 0)$ on the Argand diagram about an angle θ moves it to point $(\cos\theta, i \sin\theta)$, which amounts to a rotation by angle θ. If we now multiply this result by a scale factor s, the length of the vector from the origin to the rotated point would scale by s. Part (b) shows how point $(5, i) = 5 + 1i$ is

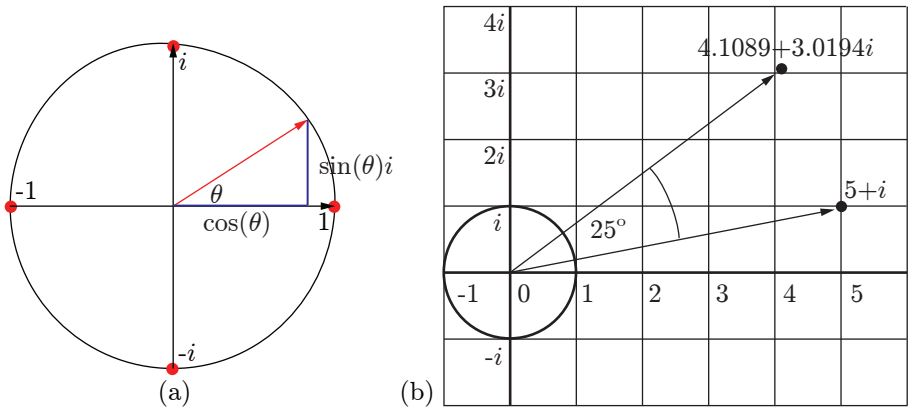


Figure 2.32: Rotation in the Argand Diagram.

rotated 25° counterclockwise by the product $\big(\cos(25^\circ)+\sin(25^\circ)i\big)(5+1i)$ which brings it to location $4.10892+3.0194i$.

Applying quaternions to three-dimensional rotations is done similarly. We choose the direction of rotation, a unit vector $\mathbf{k}=(xi,yj,zk)$ where $x^2+y^2+z^2=1$, and construct the unit quaternion

$$q(\theta/2,\mathbf{k})=\big(\cos(\theta/2)+\sin(\theta/2)\mathbf{k}\big)=\big(\cos\theta/2)+\sin(\theta/2)(xi,yj,zk)\big),$$

as the sum of a real scalar and a three-part imaginary vector $\mathbf{k}$ (relax, the reason for using half of θ is explained soon). The three-dimensional point to be rotated is denoted by $\mathbf{P}=(p_x,p_y,p_z)=(p_xi+p_yj+p_zk)$. The rotation is then preformed by

$$\begin{aligned}\mathbf{P}^* &= q(\theta/2,\mathbf{k})\,\mathbf{P}\,q^{-1}(\theta/2,\mathbf{k})\\ &= \big(\cos(\theta/2)+\sin(\theta/2)\mathbf{k}\big)\big(p_xi+p_yj+p_zk\big)\big(\cos(-\theta/2)+\sin(-\theta/2)\mathbf{k}\big)\\ &= \big(\cos(\theta/2)+\sin(\theta/2)(p_xi+p_yj+p_zk)\big)\big(p_xi+p_yj+p_zk\big)\\ &\quad\big(\cos(-\theta/2)+\sin(-\theta/2)(p_xi+p_yj+p_zk)\big).\end{aligned}$$

Note that our $\mathbf{q}$ is a unit quaternion, since $\sin^2(\theta/2)+\cos^2(\theta/2)=1$. This interesting connection between quaternions and rotations is developed in detail in [Hanson 06] (especially its page 50).

Here is why only half the rotation angle is used. Following the video of reference [PenguinMaths 20], Figure 2.33 depicts the imaginary quantities i, j, and k and their negatives as the endpoints of three perpendicular axes. Multiplying all six quantities by i on the left causes the three axes to rotate counterclockwise and has the effect of pointing the i axis out of the page and toward the viewer. Multiplying the original configuration by $-i$ on the right results in the same rotation, but flips the i axis into the page, away from the viewer. Finally, multiplying the original by $-i$ on the left rotates the axes clockwise and again points the i axis out of the page and toward the viewer.

We therefore conclude that in order to rotate all three axes counterclockwise and preserve the direction of the the i axis, we need a "sandwich" triple product where the vector $\mathbf{P}$ to be rotated is multiplied on the left by i and on the right by $-i$. This, however, causes a double rotation, which is why the two rotations are each done at half the rotation angle.

◇ **Exercise 2.37:** Prove that the triple product $\mathbf{q}\cdot[0,\mathbf{P}]\cdot\mathbf{q}^{-1}$ really performs a rotation of $\mathbf{P}$ about $\mathbf{v}$ (or $\mathbf{u}$). (Hint: Perform the multiplications and show that they produce Equation (8.15), a variation of the Rodrigues' rotation formula, Equation (8.17).)

$$\mathbf{r}^* = \vec{\mathrm{OC}} + \vec{\mathrm{CP}}^* = (\mathbf{r}\bullet\mathbf{u})\mathbf{u} + \cos\theta[\mathbf{r} - (\mathbf{r}\bullet\mathbf{u})\mathbf{u}] + \sin\theta(\mathbf{u}\times\mathbf{r}). \tag{2.16}$$

As an example of quaternion rotation, consider a 90° rotation of point $\mathbf{P} = (0,1,1)$ about the y axis. The quaternion required is $\mathbf{q} = [\cos 45°, \sin 45°(0,1,0)]$. It is a unit quaternion, so its inverse is $\mathbf{q}^{-1} = [\cos 45°, -\sin 45°(0,1,0)]$. The rotated point is thus

$$\begin{aligned}&\mathbf{q}[0,\mathbf{P}]\mathbf{q}^{-1}\\ &= [-\sin 45°, (\sin 45°, \cos 45°, \cos 45°)]\,[0,(0,1,1)]\,[\cos 45°, -\sin 45°(0,1,0)]\\ &= [0,(1,1,0)].\end{aligned}$$

The quaternion resulting from the triple product always has a zero scalar. We ignore the scalar and find that the point has been moved by the rotation from the $x = 0$ plane to the $z = 0$ plane.

Quaternion computations tend to be messy, which is why they are best performed by software. The example below is especially simple and is done by the Mathematica code:

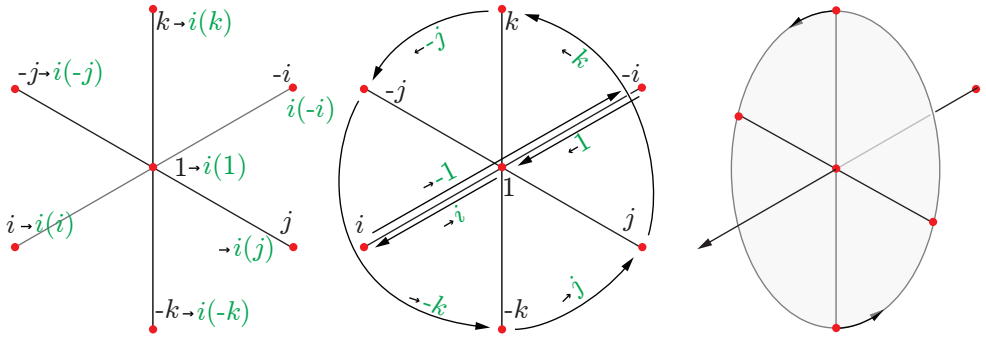
```
1 << Quaternions‘
2 q1 = Quaternion[Cos[45 Degree], 0, Sin[45 Degree], 0];
3 P = Quaternion[0, 0, 1, 1];
4 q1 ** P ** Conjugate[q1]
5 FromQuaternion[q1 ** P ** Conjugate[q1]]
```

which outputs the result, a unit quaternion, in two formats: `Quaternion[0, 1, 1, 0]` (from line 4) and `I + J` (from line 5).

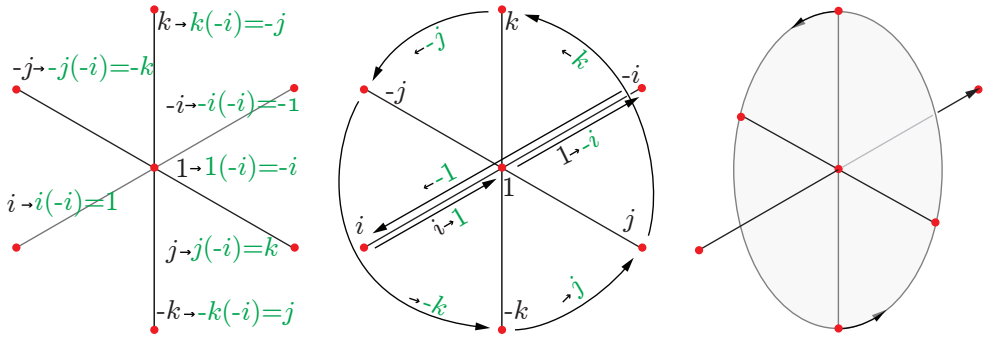
Figure 2.34 illustrates this particular rotation about the y axis and also makes it easy to understand the rule for the direction of the quaternion rotation $\mathbf{q}[0,\mathbf{P}]\mathbf{q}^{-1}$. The rule is: Let $\mathbf{q} = [s,\mathbf{v}]$ be a rotation quaternion in a right-handed three-dimensional coordinate system. To an observer looking in the direction of $\mathbf{v}$, the triple product $\mathbf{q}\cdot[0,\mathbf{P}]\cdot\mathbf{q}^{-1}$ rotates point $\mathbf{P}$ clockwise. For a negative rotation angle, the rotation is counterclockwise. In a left-handed coordinate system (Figure 2.34b), the direction of rotation is the opposite.

Euler equation for Quaternions

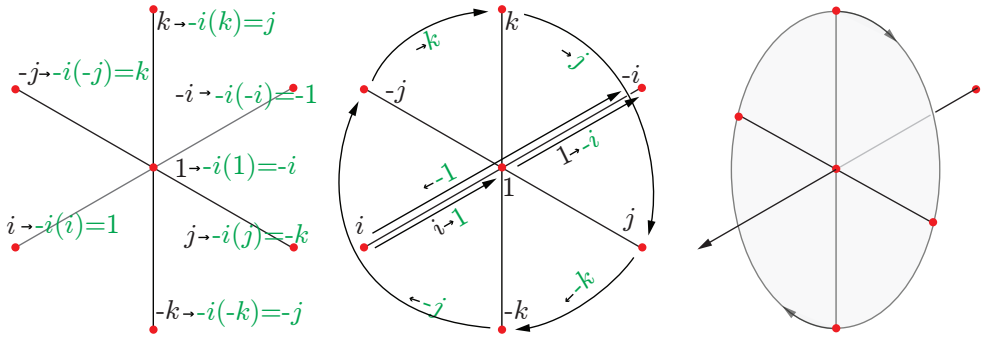
The Euler equation or identity $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ (Equation (4.6)) and its special case $e^{i\pi} + 1 = 0$ are well-known and are considered two of the most beautiful jewels of mathematics. The next few paragraphs, based on the lucid derivation at [Mathoma 15], show how to generalize this equation to e^{iq}, where q is a quaternion. Such an extension is valid and has precise mathematical meaning.



Multiply by i on the left



Multiply by -i on the right



Multiply by -i on the left

Figure 2.33: Rotating by Multiplying by i.

> . . . quaternions appear to exude an air of nineteenth century decay, as a rather unsuccessful species in the struggle-for-life of mathematical ideas. Mathematicians, admittedly, still keep a warm place in their hearts for the remarkable algebraic properties of quaternions but, alas, such enthusiasm means little to the harder-headed physical scientist.
>
> —Simon L. Altmann (1986).

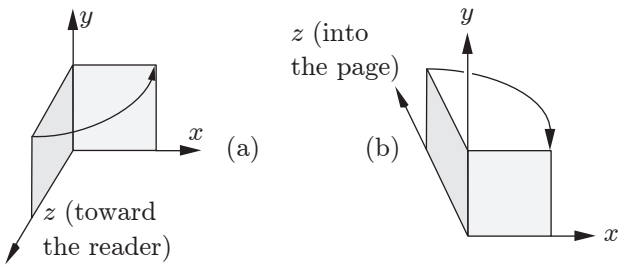


Figure 2.34: Rotation in a Right-Handed (a) and in a Left-Handed (b) Coordinate Systems.

We denote a general quaternion as the pair $(w, \mathbf{v}) = (w, a\mathbf{i} + b\mathbf{j} + c\mathbf{k})$ of a scalar w and a vector $\mathbf{v}$. The case $w = 0$ is special, and Equation (2.15) implies that the square of the quaternion $q = (0, \mathbf{v})$ is the simple $q^2 = (0, \mathbf{v})(0, \mathbf{v}) = (-\mathbf{v} \cdot \mathbf{v}, \mathbf{0}) = -|\mathbf{v}|^{\mathbf{2}}$, where $|\mathbf{v}| = \sqrt{a^2 + b^2 + c^2}$ is the absolute value of $\mathbf{v}$. Thus. For example, $q = (0, (1, 0, -2)) = i - 2k$, implying that $q^2 = -5$.

Continuing in this way, it is easy to see that $q^3 = q^2 \cdot q = -|\mathbf{v}|^2 \cdot \mathbf{v}$, $q^4 = |\mathbf{v}|^4$, and in general, even powers of our special q have only a scalar component, while its odd powers are a product of some scalar and the original vector $\mathbf{v}$.

The next step in our march toward e^{iq} is the Taylor series expansion of e^x

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots,$$

into which we substitute the powers of $q = (0, \mathbf{v})$ as follows

$$\begin{aligned} e^q &= 1 + q + \frac{q^2}{2!} + \frac{q^3}{3!} + \frac{q^4}{4!} + \cdots \\ &= 1 + \mathbf{v} - \frac{|\mathbf{v}|^2}{2!} - \frac{|\mathbf{v}|^2\mathbf{v}}{3!} + \frac{|\mathbf{v}|^4}{4!} + \frac{|\mathbf{v}|^4\mathbf{v}}{5!} - \frac{|\mathbf{v}|^6}{6!} - \frac{|\mathbf{v}|^6\mathbf{v}}{7!} + \cdots. \end{aligned}$$

Examining this expression, the first thing that comes to mind is to partition it into the even powers, which are only scalars, and the odd powers, which are multiples of $\mathbf{v}$. They are

$$\begin{aligned} e^q &= e^{ai+bj+ck} \\ &= 1 - \frac{|\mathbf{v}|^2}{2!} + \frac{|\mathbf{v}|^4}{4!} - \frac{|\mathbf{v}|^6}{6!} + \frac{|\mathbf{v}|^8}{8!} - \frac{|\mathbf{v}|^{10}}{10!} + \cdots \\ &+ \left[1 - \frac{|\mathbf{v}|^2}{3!} + \frac{|\mathbf{v}|^4}{5!} - \frac{|\mathbf{v}|^6}{7!} + \frac{|\mathbf{v}|^8}{9!} - \frac{|\mathbf{v}|^{10}}{11!} \cdots\right] \mathbf{v} \\ &= \cos(|\mathbf{v}|) + \frac{\sin(|\mathbf{v}|)}{|\mathbf{v}|}(ai + bj + ck), \end{aligned} \tag{2.17}$$

The last line of Equation (2.17) is based on the power series (Taylor series expansions

On the 16th day of [October], which happened to be a Monday, and a Council day of the Royal Irish Academy, I was walking in to attend and preside, and your mother was walking with me, along the Royal Canal, to which she had perhaps driven; and although she talked with me now and then, yet an undercurrent of thought was going on in my mind, which gave at last a result, whereof it is not too much to say that I felt at once the importance. An electric circuit seemed to close; and a spark flashed forth, the herald (as I foresaw, immediately) of many long years to come of definitely directed thought and work, by myself if spared, and at all events on the part of others, if I should even be allowed to live long enough distinctly to communicate the discovery. Nor could I resist the impulse unphilosophical as it may have been, to cut with a knife on a stone of Brougham (Broome) Bridge, as we passed it, the fundamental formula with the symbols, i, j, k; namely $i^2 = j^2 = k^2 = ijk = -1$, which contains the solution of the problem.

—Hamilton in a letter to his son Archibald (1843).



Figure 2.35: Image by MS Copilot, Sept. 2025.

around 0) of the sine and cosine functions, which are

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots$$
$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \cdots.$$

An example is $e^{i-2k} = \cos(\sqrt{5}) + \frac{\sin(\sqrt{2})}{\sqrt{2}}(i-2k) = -0.617273 + 0.351845i - 0.70369k$. Notice that Equation (2.17) is in fact a proper generalization of the original Euler

formula, because for $j = k = 0$, the equation becomes

$$e^{ix} = \cos(x) + \frac{\sin(x)}{x}(ix) = \cos(x) + i\,\sin(x).$$

We are now ready for the general case $q = (a, \mathbf{v}) = \big(a, (bi, cj, dk)\big)$, where the scalar part of the quaternion may be nonzero. This is trivial and is based on the relation

$$e^{a+bi+cj+dk} = e^a\, e^{bi+cj+dk}. \tag{2.18}$$

Equation (2.18) can now be used to generalize Equation (2.17) to the final form

$$e^{a+bi+cj+dk} = e^a \left[\cos(|\mathbf{v}|) + \frac{\sin(|\mathbf{v}|)}{|\mathbf{v}|}(bi + cj + dk)\right],$$

which makes it easy to compute the Euler form of an arbitrary quaternion. An example is

$$e^{3+i-2k} = e^3\, e^{i-2k} = e^3(-0.617273+0.351845i-0.70369k) = -12.3983+7.067i-14.134k.$$

2.9.1 Concatenating Rotations

Sometimes we have to perform two consecutive rotations on an object. This turns out to be easy and numerically stable with a quaternion representation.

If $\mathbf{q}_1$ and $\mathbf{q}_2$ are unit quaternions representing the two rotations, then associativity of quaternion multiplication implies that the combined rotation of $\mathbf{q}_1$ followed by $\mathbf{q}_2$ is represented by the quaternion $\mathbf{q}_2 \cdot \mathbf{q}_1$. The proof is

$$\mathbf{q}_2 \cdot (\mathbf{q}_1 \cdot \mathbf{P} \cdot \mathbf{q}_1^{-1}) \cdot \mathbf{q}_2^{-1} = (\mathbf{q}_2 \cdot \mathbf{q}_1) \cdot \mathbf{P} \cdot (\mathbf{q}_1^{-1} \cdot \mathbf{q}_2^{-1}) = (\mathbf{q}_2 \cdot \mathbf{q}_1) \cdot \mathbf{P} \cdot (\mathbf{q}_2 \cdot \mathbf{q}_1)^{-1}.$$

Quaternion multiplication involves fewer operations than matrix multiplication, so combining rotations by means of quaternions is faster. Performing fewer multiplications also implies better numerical accuracy.

In general, we use 4×4 transformation matrices to express three-dimensional transformations, so we would like to be able to express the rotation $\mathbf{P}^* = \mathbf{q}[0, \mathbf{P}]\mathbf{q}^{-1}$ as $\mathbf{P}^* = \mathbf{PM}$, where $\mathbf{M}$ is a 4×4 matrix. Given the two quaternions $\mathbf{q}_1 = w_1 + x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k} = (w_1, x_1, y_1, z_1)$ and $\mathbf{q}_2 = w_2 + x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k} = (w_2, x_2, y_2, z_2)$, their product is

$$\begin{aligned}\mathbf{q}_1 \cdot \mathbf{q}_2 =& (w_1w_2 - x_1x_2 - y_1y_2 - z_1z_2) + (w_1x_2 + x_1w_2 + y_1z_2 - z_1y_2)\mathbf{i} \\ &+ (w_1y_2 - x_1z_2 + y_1w_2 + z_1x_2)\mathbf{j} + (w_1z_2 + x_1y_2 - y_1x_2 + z_1w_2)\mathbf{k}.\end{aligned}$$

The first step is to realize that each term in this product depends linearly on the coefficients of $\mathbf{q}_1$. This product can therefore be expressed as

$$\mathbf{q}_1 \cdot \mathbf{q}_2 = \mathbf{q}_2 \cdot \mathbf{L}(\mathbf{q}_1) = (x_2, y_2, z_2, w_2) \begin{pmatrix} w_1 & z_1 & -y_1 & -x_1 \\ -z_1 & w_1 & x_1 & -y_1 \\ y_1 & -x_1 & w_1 & -z_1 \\ x_1 & y_1 & z_1 & w_1 \end{pmatrix}.$$

When $\mathbf{L}(\mathbf{q}_1)$ multiplies the row vector $\mathbf{q}_2$, the result is a row vector representation for $\mathbf{q}_1 \cdot \mathbf{q}_2$. Each term also depends linearly on the coefficients of $\mathbf{q}_2$, so the same product can also be expressed as

$$\mathbf{q}_1 \cdot \mathbf{q}_2 = \mathbf{q}_1 \cdot \mathbf{R}(\mathbf{q}_2) = (x_1, y_1, z_1, w_1) \begin{pmatrix} w_2 & -z_2 & y_2 & -x_2 \\ z_2 & w_2 & -x_2 & -y_2 \\ -y_2 & x_2 & w_2 & -z_2 \\ x_2 & y_2 & z_2 & w_2 \end{pmatrix}.$$

When $\mathbf{R}(\mathbf{q}_2)$ multiplies the row vector $\mathbf{q}_1$, the result is also a row vector representation for $\mathbf{q}_1 \cdot \mathbf{q}_2$.

We can now write the triple product $\mathbf{q} \cdot [0, \mathbf{P}] \cdot \mathbf{q}^{-1}$ in terms of the matrices $\mathbf{L}(\mathbf{q})$ and $\mathbf{R}(\mathbf{q})$:

$$\begin{aligned} \mathbf{q}[0, \mathbf{P}]\mathbf{q}^{-1} &= \mathbf{q}([0, \mathbf{P}] \cdot \mathbf{q}^{-1}) = \mathbf{q}([0, \mathbf{P}]\mathbf{R}(\mathbf{q}^{-1})) \\ &= ([0, \mathbf{P}]\mathbf{R}(\mathbf{q}^{-1}))\mathbf{L}(\mathbf{q}) = [0, \mathbf{P}](\mathbf{R}(\mathbf{q}^{-1})\mathbf{L}(\mathbf{q})) \\ &= [0, \mathbf{P}]\mathbf{M}, \end{aligned}$$

where matrix $\mathbf{M}$ is

$$\begin{aligned} \mathbf{M} &= \mathbf{R}(\mathbf{q}^{-1}) \cdot \mathbf{L}(\mathbf{q}) \\ &= \begin{pmatrix} w & z & -y & x \\ -z & w & x & y \\ y & -x & w & z \\ -x & -y & -z & w \end{pmatrix} \begin{pmatrix} w & z & -y & -x \\ -z & w & x & -y \\ y & -x & w & -z \\ x & y & z & w \end{pmatrix} \\ &= \begin{pmatrix} w^2+x^2-y^2-z^2 & 2xy+2wz & 2xz-2wy & 0 \\ 2xy-2wz & w^2-x^2+y^2-z^2 & 2yz+2wx & 0 \\ 2xz+2wy & 2yz-2wx & w^2-x^2-y^2+z^2 & 0 \\ 0 & 0 & 0 & w^2+x^2+y^2+z^2 \end{pmatrix}. \end{aligned}$$

Since we have unit quaternions, they satisfy $w^2 + x^2 + y^2 + z^2 = 1$, so we can write the final result

$$\mathbf{M} = \begin{pmatrix} 1-2y^2-2z^2 & 2xy+2wz & 2xz-2wy & 0 \\ 2xy-2wz & 1-2x^2-2z^2 & 2yz-2wx & 0 \\ 2xz+2wy & 2yz-2wx & 1-2x^2-2y^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{2.19}$$

In a left-handed coordinate system, the same rotation is expressed by the triple product $\mathbf{q}^{-1}[0, \mathbf{P}]\mathbf{q}$ or, equivalently, by $\mathbf{P}^* = \mathbf{P} \cdot \mathbf{M}^T$, where $\mathbf{M}^T$ is the transpose of $\mathbf{M}$.

My name is William Rowan Hamilton and no one uses my quaternion.

2.10 Transcendental Numbers

Definition: A transcendental number is a real or complex number that is not algebraic, that is, not the root of a non-zero polynomial of finite degree with rational coefficients. The word "transcendental" comes from the Latin and means to climb over or beyond, to surmount.

Transcendental numbers

"Transcendental numbers, like π and e, lie beyond the algebraic; they defy the solutions of polynomial equations, transcending the simple roots and revealing the infinite complexities of the mathematical universe."
This quote encapsulates the essence of transcendental numbers, highlighting their unique and complex nature in the realm of mathematics.

—By `chatGPT 4o`, May 2024.

Every real number is either algebraic (a solution of an algebraic equation) or transcendental (not a solution of any such equation). However, very few transcendental numbers are known (among them are π and e), which is why in the past, mathematicians tended to assume that they are rare. Unexpectedly, in 1874, while developing his set theory, Georg Cantor managed to prove that the transcendentals vastly outnumber the algebraic numbers. Cantor did not discover any transcendentals, he only proved that they exist, and yet this existence proof is a surprising result that is equivalent to the following statement. If you select a real number at random among all the real numbers, it will be transcendental with probability 1!

In 1844 Joseph Liouville proved that transcendental numbers exist, and in 1851 he gave the first examples of such numbers, among them the Liouville constant

$$L_b = \sum_{n=1}^{\infty} 10^{-n!}.$$

This strange number starts with "0." followed mostly by zeros, where the nth digit is 1 only if n is the factorial of some positive integer.

Table 2.36 lists a few algebraic equations and their solutions. It is clear that the solutions can be integers, rational, irrational, and even complex numbers, but all these solutions are algebraic, and therefore not transcendental.

Equation	Solution
$x - 5 = 0$	5
$5x - 4 = 0$	$4/5$
$x^2 - 2 = 0$	$\sqrt{2}$
$x^2 + 1 = 0$	i
$x^4 - 22x^2 + 1 = 0$	$\sqrt{11 + 2\sqrt{30}}$

Table 2.36: Several Algebraic Numbers.

Transcendental numbers are different. Consider, for example, e. This number can be represented by the infinite sum

$$e = \sum_{k=0}^{\infty} \frac{1}{k!},$$

all of whose terms are algebraic, but because of the mysterious, unpredictable nature of the infinite, it was possible for Charles Hermite to prove, in 1873, that no algebraic equation yields e as a solution.

Ferdinand von Lindemann proved the transcendence of π in 1882. He first showed that e^a is transcendental if $a \neq 0$ is an algebraic number. Then concluded that $i\pi$ must be transcendental because $e^{i\pi} = -1$ is algebraic. But since i is algebraic, π must be transcendental. As mentioned on Page 8, the transcendence of π implies that the classical compass and straightedge geometric constructions of the ancient Greeks cannot include transcendental numbers.

> Die Mathematiker sind eine Art Franzosen; redet man mit ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsobald ganz etwas anderes.
> (Mathematicians are [like] a sort of Frenchmen; if you talk to them, they translate it into their own language, and then it is immediately something quite different.)
> —Johann Wolfgang von Goethe.

2.11 Important and Interesting Numbers

Some real numbers have unusual, unexpected, and interesting or useful properties. The best known ones are π, e, φ (the golden ratio), $i = \sqrt{-1}$, and zero. The next subsections discuss those numbers and their applications.

2.11.1 Euler's number e

The Euler number e is one of the important mathematical constants—the numbers π, φ (the golden ratio), $i = \sqrt{-1}$, and zero—that seem to pop out of nowhere while also appearing everywhere. Notice that there is also an Euler's constant, denoted by γ, which is an extension of the well-known factorial.

Euler discovered the number e by studying the behavior of continuous compound interest. In particular, he was interested what happens when interest on an investment is continually compounded over time. He found that the rate of growth of a continuously compounded investment is equal to the number e raised to the power of the interest rate. This is known as Euler's formula, and it is one of the most important results in mathematics. Euler's work on this topic was published in a paper called "Institutiones calculi integralis" in 1748.

Text by chatGPT, June 2023.

Many references claim that although commonly associated with Leonhard Euler (pronounced "Oiler") and named after him, e was first discovered, in 1683, by Jacob Bernoulli as part of his research on how wealth grows when interest rates are compounded often, instead of annually.

Over the years, it has been discovered that e appears in many mathematical and physical contexts. This number is also irrational (it cannot be represented as a ratio of integers), transcendental (it is not a root of any non-zero polynomial with rational coefficients), and is the base of the Natural Logarithms (introduced by John Napier).

The following sentence is a mnemonic that helps memorize some digits of e: "to express e remember to memorize a sentence to memorize this." The lengths of the words of this sentence are 2, 7, 1, 8, 2, 8, 1, 8, 2, 8, and 4, so they give $e \approx 2.7182818284$ to an accuracy of 11 decimal digits.

There are various ways to introduce e, and here it is discussed and defined as a special exponential function. A function $f(x)$ of the form b^x, where b is a real constant, is referred to as exponential. Exponential functions are proportional to their derivatives, which is why it makes sense to search for the base b of the special exponential function b^x whose proportionality constant is 1. Such a function, denoted by $exp(x)$, has the unique property that it equals its derivative.

The following paragraph derives this function from the condition that it be exponential and will equal its derivative. Once we find the base b of this special b^x function, we name it e and try to compute its numeric value.

We start with the definition of the derivative of b^x

$$\begin{aligned}\frac{d}{dx}[b^x] &= \lim_{\Delta x\to 0}\frac{b^{x+\Delta x}-b^x}{\Delta x} = \lim_{\Delta x\to 0}\frac{b^x b^{\Delta x}-b^x}{\Delta x}\\ &= b^x \lim_{\Delta x\to 0}\frac{b^{\Delta x}-1}{\Delta x}.\end{aligned}$$

We are almost there! The derivative of b^x is now expressed as itself times a strange, annoying limit that we must coerce to become 1. This turns out to be quite easy. From the condition $(b^{\Delta x}-1)/\Delta x = 1$ we derive $b = (1+\Delta x)^{\frac{1}{\Delta x}}$. Finally, we rename b and Δx, e and $1/n$, respectively, to end up with

$$e = \lim_{n\to\infty}(1+\frac{1}{n})^n.$$

The above expression is often considered the definition of e. Mathematical rigor requires that we prove that the limit exists. Numerical computations, especially easy today, when fast mathematics software is available, produces its numeric value. Other ways to look at e are (1) It is the unique positive number a such as the exponential function a^x has a slope of 1 at $x = 0$. (2) It is the sum of the infinite series

$$\sum_{n=0}^{\infty}\frac{1}{n!} = 1+\frac{1}{1}+\frac{1}{1\cdot 2}+\frac{1}{1\cdot 2\cdot 3}+\cdots.$$

The natural logarithm $\ln = \log_e$ (logarithm to base e) is the inverse of $\exp(x)$, so

the relation between and e and ln is

$$\ln e \stackrel{\text{def}}{=} \int_1^x \frac{1}{t} dt = 1.$$

The origin of the term "natural logarithm."

In calculus, the general expression for a derivative is the limit

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

When this is applied to computing the derivative of the logarithm, $\log_b(x)$, the corresponding expression is the limit

$$\lim_{h\to 0} \frac{\log_b(x+h) - \log_b(x)}{h},$$

which can be manipulated to become

$$\frac{1}{x}\log_b \left[\lim_{h\to 0}\left(1+\frac{h}{x}\right)^{x/h}\right] = \frac{1}{x}\log_b\, e.$$

and it is this expression of the derivative that makes it *natural* to choose the Euler constant e as its base b, and to refer to this base as natural.

◇ **Exercise 2.38:** The following is actually a simple problem. What function equals the reciprocal of its derivative?

Following are a few interesting and unexpected properties of and appearances of e in mathematics and science.

- To understand the meaning of "e^x equals its derivative," imagine a car accelerating on a racing course in such a way that its position, velocity, and acceleration remain the same as it races. When it has covered 13 km, its speed is 13 kph and its acceleration is 13 kph-per-hour.
- Here is a mathematical game where e pops up unexpectedly. Given a real number a, Choose an integer n, divide $b = a/n$, and multiply $r = b^n = (a/n)^n$. The product r is a real number that depends on both a and n. It turns out that r is at its maximum when $b = a/n$ is closest to e. The Mathematica code `Solve[D[(a/n)^n, n]==0.]` illustrates why this is so, although this code is not a formal proof. The code first differentiates the function $(a/n)^n$ with respect to n and then finds the value of n where the derivative is zero. A zero derivative of a smooth function implies a maximum or minimum of the function. Executing this code produces $n- > 0.367879a$, very close to $(1/e)a$.
- Draw several random numbers in the interval $[0, 1]$ and add them. Keep drawing and adding until the sum exceeds 1. Assuming that this took n numbers, it can be shown that the expected value (the average) of n is e.
- Perhaps the easiest way to understand e is to consider its appearance in computations of interest rates. For simplicity, we assume a \$1 investment that earns 100% year. If

the interest is paid at the end of the year, the final payout is $1 + 1 = 2$. However, if the interest is paid every six months and is immediately reinvested, then after the first six months the investment is worth 1.50 which after the next six months earns 1.75, for a total payout of $1 + 0.50 + 0.75 = 2.25$.

Similarly, if the interest is paid every three months, the four 25% interest payments would be 0.25, 0.325, 0.3906, and 0.4882, and the investment would grow during the year from 1 to 1.25, 1.5625, 1.9531, and finally 2.4413.

It seems that the total payout increases as the interest is paid more often, but the increases themselves get smaller and may prove to converge to a finite limit. Thus, the obvious question is what happens at the limit, when the interest is paid continually. This cannot be done in practice, as there is always a smallest unit of currency, but it can be computed theoretically and it turns out to be a payout of of $e = 2.718$.

- The next example where e appears is in probability computations. Imagine a lottery where the chance of winning is 10^{-6} (one in a million). Naturally, the chance of a loss in each play is the much higher 0.999999 or

$$\left(1 - \frac{1}{1000000}\right).$$

Playing again and again, the chance of winning or losing each time are the same 10^{-6} and 0.999999, respectively, but the chance of losing *every* time gets smaller with more plays. This is because individual plays are independent events, so the chance of losing every time is the product of the individual chances. For n plays, this chance is

$$\left(1 - \frac{1}{1000000}\right)^n$$

And the limit of this quantity when n approaches infinity is $1/e$ or approximately 36.79%).

- Consider the unusual infinite equation

$$x^{x^{x^{\cdot^{\cdot^{\cdot}}}}} = 2.$$

This scary tetration is easy to solve when we write it as $x^y = 2$ and then realize that y, being an infinite ladder of x's, equals the entire left-hand side of the equation and therefore also equals the right-hand side, which is 2. We can then rewrite the equation as $x^2 = 2$ whose solution is $x = \sqrt{2}$. We have just proved that

$$\sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\cdot^{\cdot^{\cdot}}}}} = 2.$$

Now consider the similar equation

$$x^{x^{x^{\cdot^{\cdot^{\cdot}}}}} = 4.$$

We can similarly rewrite it as $x^y = 4$ or $x^4 = 4$, with a solution $x^2 = 2$ and $x = \sqrt{2}$. We have now proved that

$$\sqrt{2}^{\sqrt{2}^{\sqrt{2}^{\cdot^{\cdot}}}} = 4.$$

The conclusion is that either $2 = 4$ or that we made a mistake somewhere. The mistake was to treat infinity naively, without the respect it deserves. It turns out that $\sqrt{2}$ is a solution to the first equation, but the second equation has no solutions because the solutions of equations of the form

$$x^{x^{x^{\cdot^{\cdot}}}} = \text{constant},$$

converge only if $x \in [e^{-e}, e^{1/e}] = [0.06599, 1.4447]$ and this interval contains $\sqrt{2}$ but not $\sqrt{4}$. This unexpected result was also proved by Euler. See reference 23 in [WikiTetrat 23].

Other appearances of e in common situations are:

- Take a hot dish out of the oven and let it cool down naturally. The graph of its dropping temperature as a function of time is an exponential of the form $C + ae^{kt}$ where the initial temperature is $A + C$, the final temperature is C, and k is a constant that depends on the material of the dish.
- Imagine a weight hanging at the end of a spring. When pulled down in a uniform gravity field and let go, the weight will oscillate up and down, losing amplitude continually. The curve of its amplitude as a function of time has an envelope (Figure 2.37) that is given by an exponential expression of the form ae^{-kt}. It drops from an initial amplitude of a to zero.

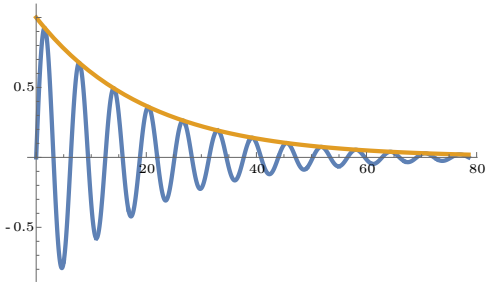


Figure 2.37: Exponential Envelope of a Damped Oscillator.

> Gentlemen, that is surely true, it is absolutely paradoxical; we cannot understand it, and we don't know what it means. But we have proved it, and therefore we know that it is the truth.
>
> —Benjamin Peirce.

2.11.2 π, Not Just a Circle

People like the title "king." We often hear declarations such as He is the king of washing machines or the king of bail bonds. Similarly, if you ask someone about the king of numbers, a likely answer would be π. The number π is likely to show up whenever we study mathematics, engineering, science, or any other field where numbers are used. It is hard to believe that π, which is defined in relation to circles, is such an important part of our lives, yet no one has been able to explain the universality and importance of this number.

As most everyone knows, π is defined as the ratio of the circumference of a circle to its diameter, and its approximate value is 3.1415. Already the ancients have noticed that this ratio is constant and applies to all circles, much as the trigonometric ratios sin and cos apply to all right-angle triangles. We can check this claim by computing π for various circles, but we cannot explain it or prove it formally. The interested reader can try the simple Mathematica code `Integrate[r/Sqrt[r^2-x^2],{x,-r,r}]/r` for various values of r. It is perhaps this generality of π that makes it so universal, but we can only agree that this property of the circle is one of the mysteries of mathematics.

Already the Babylonians and Egyptians tried to compute the value of π and came up with numbers close to 3, but the first serious attempt in this direction was made by Archimedes. He imagined two regular polygons (polygon: a plane figure bounded with straight segments) one circumscribing the circle from the outside and the other inscribing the circle from the inside. The areas of the polygons are relatively easy to calculate, and the area of the circle is somewhere between them. By gradually increasing the number of polygon sides, Archimedes was able to obtain better and better approximations of π.

◇ **Exercise 2.39:** Find several approximations of π made in the past.

Over the centuries, many mathematicians and mathematical fans discovered other techniques to compute π and came up with ridiculously long approximations, but it was only in 1768 that Johann Lambert managed to prove that π is irrational and thus can never be fully expressed as a decimal number. Finally, in 1882 Ferdinand von Lindemann proved that e^a is transcendental for every non-zero algebraic number a, thereby establishing that π is transcendental (i.e., it is not the solution of any non-constant polynomial equation with rational coefficients).

The transcendence of a number is much more difficult to prove than irrationality, and has a number of mathematically important consequences, one of which is that a transcendental number cannot be constructed with compass and straightedge. This is why the transcendence of π implies that it is impossible to square the circle, an important goal of ancient geometers. (This is also why the term "circle squarer" is derogatory, but see story on Page 8.)

A curiosity. In 1897, an American circle squarer has almost succeeded in persuading the Indiana legislature to pass a Bill that defined π as 3.2, because this value resulted from his circle squaring method. Luckily for humanity, the bill never became a law (imagine the response of visiting aliens to this.).

The Greek letter π (pi) was adopted, by William Jones in 1706, to denote this important number.

The well-known Euler's formula $e^{i\pi} - 1 = 0$ contains π among the other important constants e, i, 1, and 0.

> We stared at the note. No one moved. On the paper he had written a single line, one simple formula:
>
> $$e^{i\pi} + 1 = 0.$$
>
> No one spoke. The widow's fingernails had ceased their tapping. Her eyes, so full of suspicion and disdain a moment earlier, now looked at me with a calm, understanding gaze, and I could tell then that she knew the beauty of math.
>
> —Yōko Ogawa, *The Housekeeper and the Professor*, (2003).

Pi Day

It was Larry Shaw, a physicist, tinkerer, curator, and media specialist at the San Francisco Exploratorium, who became the Prince of Pi when in 1988 he linked March 14 (3.14) first with the digits of π and later also with Einstein's birthday (in 1879).

Since 1988, many math lovers, educators, and pi enthusiasts the world over celebrate this day and even raise a slice of pie (a π pie) precisely at 1:59 (more π digits) to proclaim the importance of the king. In March 2009, Pi Day became an official U.S. national holiday. Common activities on Pi Day include a parade at the exploratorium.

Observing that the fraction 22/7 is a common approximation of π, an anonymous enthusiast, deciding not to lag behind Larry Shaw, came up with Pi Approximation Day, which is observed on July 22 (22/7 in the day/month date format).

Pi Day in 2015 was so far the most important one, because at 9:26:53 the time was 3|14|15|9|26|53, reflecting the first 10 digits of π.

◇ **Exercise 2.40:** (Fun). Search for other celebrities who were born or died on Pi Day.

The Pi Day website is at `https://www.piday.org/`

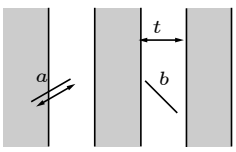
A coincidence? Denote the thickness of a pizza by a and its radius by z. Its volume would be $\pi \cdot z^2 \cdot a = Pizza$ (by Presh Talwalkar).

Approximations of π

Because of the importance and universality of π, generations of mathematicians have searched for ways to easily compute many of its digits. Most of these methods employ infinite series, but here we present two ideas that approach this problem from different directions. The first idea is to use probabilities, and the simplest algorithm in this family is the so-called Monte Carlo method, which simply counts the results of multiple random trials.

Imagine a radius 1 circle inscribed in a unit square. A computer program draws many pairs (a, b) of random numbers, each in the interval $[0, 1]$. Naturally, each such pair can be considered a point that is located in or on the square, but certain pairs also end up in or on the circle. Those are pairs that satisfy $\sqrt{a^2 + b^2} \leq 1$. Simply count the number m of such pairs and divide by the total number n of pairs generated and checked. The ratio m/n should approach the ratio of the area of the square (which is 1) to the circle's area (which is $\pi(1^2)$), as more and more pairs are computed and tested.

Another probability-based algorithm for π is the Buffon needle, illustrated in the figure. Parallel lines are drawn on a surface,



t units apart. A needle of length l is dropped many times at random on the surface. Sometimes, the needle end up crossing a line, and those cases are denoted by a. In other cases, b in the figure, the needle ends up between lines. it can be shown that the ratio $2(a+b)l/(at)$ is an approximation of π, which gets better as the number of drops gets bigger.

The next approach to approximate π is the spigot algorithms, which first appeared in 1995. Those algorithms produce individual digits of π where each is independent of its predecessors. Thus, there is no need to save newly obtained digits for future use. This is analogous to water droplets dripping from a spigot. These algorithms are too long and involved to be described here, but here are a few words about each.

The first spigot algorithm was due to Stanley Rabinowitz and Stan Wagon. See reference [Rabinowitz 95].

The Bailey-Borwein-Plouffe (BBP) algorithm, reference [WikiBBP 95], is sometimes referred to as magical because it can produce any hexadecimal digit of π without having to compute its predecessors. It is based on the series

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right).$$

The following mnemonic text, due to James Jeans, generates the first 24 digits of π, each is the length of a word in this text.

"How I want a drink, alcoholic of course, after the heavy chapters involving quantum mechanics. All of thy geometry, Herr Planck, is fairly hard ..."

The Gauss-Legendre formula; fast computation of π.

One of the many contributions of Gauss to science and mathematics is a formula for computing π to any precision. Even though this formula is iterative and complex-looking, it is useful because it converges rapidly. A precision of about 20 digits is typically achieved after only three iterations! The main component of this formula is the AGM function, so called because it produces the arithmetic-geometric mean of its two parameters. Section 10.8 discusses arithmetic average and geometric mean, two concepts that are employed by AGM.

The principle of AGM is simple. Start with two real values a_0 and b_0, and iterate each according to

$$a_{k+1} = \frac{a_k + b_k}{2} \quad \text{and} \quad b_k = \sqrt{a_k b_k}.$$

In each iteration, a_k and b_k move toward each other. In fact, the number of leading digits where a_k and b_k *do not* differ, approximately doubles in each iteration, which is why AGM features quadratic convergence. In his original work from around 1800, Gauss started with $a_0 = 1$ and $b_0 = 1/\sqrt{2}$ and after only three iterations ended up with a_k and b_k that agreed in their 19 leading digits.

Once the principle of AGM is clear, the Gauss-Legendre formula for π can be written

$$\pi = \frac{2\text{AGM}^2(1, 1/\sqrt{2})}{\frac{1}{2} - \sum_{k=1}^{\infty} 2^k c_k^2},$$

where $c_{k+1} = (a_k - b_k)/2$ and $c_{k+1}^2 = a_{k+1}^2 - b_{k+1}^2 = (a_{k+1} - a_k)^2$. The particular values $a_0 = 1$ and $b_0 = 1/\sqrt{2}$ were chosen by Gauss, but may not be the only ones for which this formula works and produces approximations of π.

The formula seems complex. It requires iterations both in the AGM function and in its denominator, but it is still useful because of its rapid convergence. Its first three iterations produce

$$\begin{aligned}
&3.140\ldots\\
&3.14159264\ldots\\
&3.1415926535897932382\ldots \qquad .\\
&3.14159265358979323846264338327\ldots
\end{aligned}$$

Chapter 7 of reference [Arndt 06] has more about this formula, its properties and history.

Piphilology

The constant π has fired The imagination of people for millennia. One aspect of fascination is its digits. They seem random, but they are not, because they are determined uniquely by any algorithm that computes π. They are infinite and feature no pattern. They may be normal, but as of 2025 this question hasn't been resolved. Given all this, it is no wonder that some people try to achieve fame by memorizing many digits of this number.

There are many videos that purport to teach methods for memorizing more and more digits of π. Around 2015, the champion piphilologer was able (as verified by the Guinness World Records) to recite by heart the first 70,000 digits of π in about 10 hours; an empty and futile achievement. (Piphilology is the practice of memorizing large numbers of digits of π.)

One of the largest dimensions in the solar system is the distance of Voyager 1 from Earth, which in March 2025 was about 25 billion kilometers. In units of millimeter this is $25\times10^9\times10^6$, a 17-digit number. Assuming that this number is 12345678901234567 mm and that it is the radius of a circle, we compute the circumference of the circle. Using π to 15-digit accuracy, this comes to

$$\begin{aligned}
&12345678901234567\times 2\times 3.141592653589793\\
&\quad = 77,570,188,279,394,046.626064179949262\,\text{mm}.
\end{aligned}$$

Adding one more digit of precision to π changes this computation to

$$\begin{aligned}
&12345678901234567\times 2\times 3.1415926535897932\\
&\quad = 77,570,188,279,394,051.5643357404430888\,\text{mm}.
\end{aligned}$$

The difference is only five millimeters.

Today, we know that an accuracy of 39 digits is enough to calculate the circumference of the observable universe with a precision of one atom. Engineers often claim that a precision of 5–6 digits in the value of π is enough for most practical calculations,

which is why computing π to a vast number of digits is rarely useful, but may serve in the search for solving the question of whether π is normal.

A beautiful, intriguing, but illusory relation

Let's start with the simple expression $\pi^4 + \pi^5 = e^6$ (this is not an equation; there is no unknown). It shows an intriguing, hitherto unknown relation between two important mathematical constants, so we decide to look deeper into it. Noticing that each of the three terms is an exponent, we rewrite this expression as the sum of squares $(\pi^2)^2 + (\pi^{2.5})^2 = (e^3)^2$. It now looks like the Pythagorean theorem for a right-angle triangle, so we check to see how true it really is.

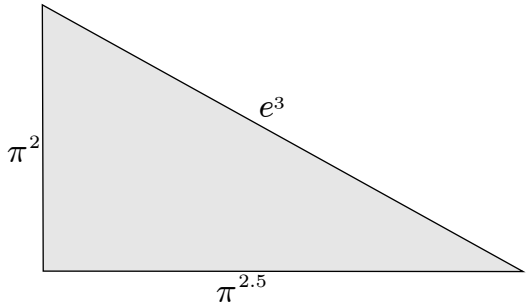


Figure 2.38: A Mock Right Triangle.

Using drawing software to draw a triangle with the above dimensions, Figure 2.38 shows what looks like a true right triangle, but the precise calculations listed here, using the cosine rule, prove that the angle between its "perpendicular" sides is slightly different from 90°, thereby dashing our hopes for newly discovered mathematical beauty, and allowing us to label this triangle as deceptive, deceiving, fallacious, or mock.

The close cousin of the well-known cosine rule yields the angle γ between sides a and b of a triangle if all three sides are known

$$\gamma = \arccos\left[\frac{a^2+b^2-c^2}{2ab}\right], \text{ applied here, } \gamma = \arccos\left[\frac{\pi^4+\pi^5-e^6}{2\pi^{4.5}}\right],$$

produces $\gamma = 1.57079637798$, which in turn implies that the "right" angle in Figure 2.38 equals $180\gamma/\pi = 90.0000029325°$.

This curious result is due to youtuber `polymathematic`.

> "An interesting point," Holmes said dreamily. "I subscribe to several esoteric journals to ease the tedium of the apiarist's life. I was reading in one of them that William Shanks, a mathematician of our fair island, has recently computed π to 707 decimal places. It took him twenty years. His result filled a whole page with quite senseless, random numbers. Should anyone doubt Mr. Shanks's result, he would have to budget an equal amount of time and duplicate his work. In that case also, verifying the answer would be precisely as difficult as coming up with the answer in the first place—the very antithesis of an 'obvious' solution."
> (Unknown to Holmes's generation, Shanks made an error on the 528th decimal place. All the digits after that were wrong.)
>
> —William Poundstone, *Labyrinths of Reason*, 2011.

Here is an unusual property of π. The definite integral
`Integrate[(x^4 (1 - x)^4)/(1 + x^2), {x, 0, 1}]` equals $(22/7) - \pi \approx 3.142$.

History tells us that Archimedes was able to calculate a very good approximation of π by constructing regular polygons both inside and outside a circle and computing their perimeters. He started with two regular hexagons, one circumscribing and the other inscribing, the circle. He then gradually increased the number of polygon sides up to 96, where he was able to estimate π to be between $3\frac{1}{7}$ and $3\frac{10}{71}$ or approximately 3.1429 and 3.1408. Such a technique is referred to as a method of exhaustion, and Archimedes applied it also to prove that the area of a circle is πr^2.

In reference [MindYourDecisions 25] YouTube contributor `MindYourDecisions` explains how the excellent approximation 355/113 of π (Table 2.21) was obtained in 480 by Zu Chongzhi, based on an earlier method developed by Liu Hui in 263. This method is similar to that of Archimedes, but computes the areas of the successive polygons instead of their perimeters.

Today, such calculations can easily and rapidly done by software, and reference [Brain Station Advanced 25] shows how to do it with a spreadsheet.

[Arndt 06], [Berggren 97], [MathsNeedNotApply 25], and [Posamentier 04] are some of the many thousands of references that discuss, explain, and explore the "mysteries" of π.

Why does π appear everywhere? Because it is defined by the properties of the circle. This definition implies that π is related to circles and therefore also to spheres and wave motion, concepts that are at the foundations of many phenomena and objects in our universe, from atomic structure to galactic scales. There are innumerable examples of π appearing in mathematical results, relations, and theorems, but the following are a few surprising examples of π showing up in non-mathematical connections:

- Hiding a message in the infinitely many digits of π. This is mentioned in the 1985 novel *Contact* by Carl Sagan.

> Whoever makes the universe hides messages in transcendental numbers so they'll be read fifteen billion years later when intelligent life finally evolves.
>
> —Carl Sagan, *Contact*, 1985.

- Donald Knuth started working on his popular TEX software in 1978. It was about 10 years later, when TEX was in good shape and started producing reasonable results,

that Knuth noticed that he was in version 3.14 of the software. He then decided to number future versions by the consecutive digits of π. The current version of TEX (last updated in 2021) is 3.141592653.

■ *Aerial* is the eighth album by the well-known singer-songwriter Kate Bush. It was released in November 2005, and climbed the charts fast.

The second track, "π," of *Aerial*, reference [Aerial 05], describes a man who has "a complete infatuation with the calculation of π." The digits of π are sung, with some errors, from the first to the 78th, then from the 101st to the 137th. They later became sequence A112602 in the OEIS.

> A112602. Erroneous version of decimal expansion of Pi (see A000796 for the correct version).
> 3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5, 8, 9, 7, 9, 3, 2, 3, 8, 4, 6, 2, 6, 4, 3, 3, 8, 3, 2, 7, 9, 5, 0, 2, 8, 8, 4, 1, 9, 7, 1, 6, 9, 3, 9, 9, 3, 7, 5, 1, 0, 5, 8, 2, 0, 9, 7, 4, 9, 4, 4, 5, 9, 2, 3, 0, 7, 8, 1, 6, 4, 0, 6, 2, 8, 6, 2, 0, 8, 8, 2, 1, 4, 8, 0, 8, 6, 5, 1, 3, 2, 8, 2, 3, 0, 6, 6, 4, 7, 0, 9, 3, 8, 4, 4, 6, 0, 9, 5, 5, 5, 2, 2, 3
>
> —From `OEIS`.

■ The videos at `t.ly/rnTh3` and `t.ly/h3PM6`, by YouTuber `3Blue1Brown`, illustrate surprising connections between π and certain physical phenomena.

■ The sinuosity index (or meandering ratio) of a stream or a river is the ratio of the length of the stream to the straight line distance between its source and mouth. This is the result of the tendency of a river to cut an easy meandering path in the ground, rather than proceed in a possibly difficult straight line. Hydrologists employ this fact to define the terms sinuous river (when the sinuosity index is in $[1, 1.5)$) and meandering river (when the sinuosity index is in $[1.5, 4]$).

There is a scientific study, summarized in reference [Stolum 96], which claims that the average of the sinuosity index of earth rivers tends to oscillate between a low of 2.7 and a high of 3.5, with an average of 3.14. However, there are many objections to this conclusion.

■ The well-known Heisenberg's Uncertainty Principle states that there is an inherent uncertainty built into the universe that becomes apparent when measuring certain properties of quantum particles. (It is too small to measure for everyday-size objects.) A common example of this principle has to do with the position and momentum of a particle, where the principle states that the more precisely the position is known the more uncertain the momentum is, and vice versa.

The principle is expressed mathematically in the form

$$\Delta x \Delta y \geq \frac{h}{4\pi} \stackrel{\text{def}}{=} \hbar.$$

The product of the uncertainty in a quantity x and the uncertainty in a quantity y is never less than the Planck's constant h divided by 4π, a constant known as h-bar.

◇ **Exercise 2.41:** Since the expansion of π is infinite, can it appear within itself? Can the infinite sequence of π digits appear as a subsequence of itself?

◇ **Exercise 2.42:** (Strange.) What if π were an algebraic number (non-transcendental)?

> Probably no symbol in mathematics has evoked as much mystery, romanticism, misconception and human interest as the number π.
>
> —William L. Schaaf, *Nature and History of Pi*.

2.11.3 i, the $\sqrt{-1}$

A reminder. How to multiply square roots of negative numbers? We know that in a mathematical expression, with numbers, letters, and arithmetic operations, the rules of operator precedence are: functions (including square roots), exponentiation, multiplication/division, and addition/subtraction. But what about square roots of negative numbers? With positive numbers, we have the simple identity $\sqrt{a \cdot b} = \sqrt{a}\sqrt{b}$, and this can be applied to negative numbers by $\sqrt{-10} = \sqrt{(-1) \cdot 10} = \sqrt{-1}\sqrt{10} = i\sqrt{10}$. Thus, for example, $\sqrt{-16} \cdot \sqrt{-2} = i \cdot \sqrt{16} \times i \cdot \sqrt{2} = -\sqrt{32}$.

The notation i for $\sqrt{-1}$ was originally introduced (by Gauss in 1831) to remove ambiguities in complex number arithmetic, such as in $\sqrt{-1}^2$ which prompts one to write it as $\sqrt{-1}\sqrt{-1}$, and thus obtain the wrong result -1.

Another look at i. We normally think of i as $\sqrt{-1}$, but Euler's formula provides another way to express this important constant. We start with $e^{iZ} = \cos(Z) + i\sin(Z)$ for an arbitrary complex number Z, and add this to its negative $e^{i(-Z)} = \cos(Z) - i\sin(Z)$ to obtain $e^{iZ} + e^{i(-Z)} = 2\cos(Z)$. When this is multiplied by e^{iZ} on both sides, it becomes the quadratic equation $(e^{iZ})^2 - 2\cos(Z)e^{iZ} + 1 = 0$, whose solution is

$$e^{iZ} = \frac{-(-2\cos(Z)) \pm \sqrt{(-2\cos(Z))^2 - 4 \cdot 1 \cdot 1}}{2 \cdot 1} = \frac{2\cos(Z) \pm 2\sqrt{\cos^2(Z) - 1}}{2},$$

which produces the new expression $i = \frac{1}{Z}\ln\bigl(\cos(Z) \pm \sqrt{\cos^2(Z) - 1}\bigr)$. The familiar constant i is now expressed in terms of an arbitrary complex number Z.

Note. I tried this with Mathematica, but it seems to depend highly on the internal precision of Z and seems to work only for complex numbers Z where both the real and imaginary parts are on the order of up to 3.

```
Z = 2 + 3 I;
a = Cos[Z]; b = a + Sqrt[a^2 - 1];
(1./ Z) Log[b]
```

And now, for a few elegant, simple examples of operations that involve i. They all follow the excellent videos of `youtube` channel `blackpenredpen`.

• What is the $\sqrt{i}$? Arithmetic operations on complex numbers normally result in complex numbers, but sometimes such a result may be real. As a first guess, we assume that $\sqrt{i}$ is complex, and start with the general expression $\sqrt{i} = a + bi$, with a and b to be determined. Not knowing what else to do with this, we square both sides, to obtain $i = a^2 + 2abi - b^2$. Since i is imaginary, its real part is zero and its imaginary part is i. We can therefore write $0 + i = a^2 + 2abi - b^2$, which implies $a^2 - b^2 = 0$ and $2ab = 1$. These two equations are easily solved to produce the final result

$$\sqrt{i} = \pm\frac{1}{\sqrt{2}}(1 + i) = \pm\left[\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i\right]. \tag{2.20}$$

This technique of deriving $\sqrt{(i)}$ can be termed algebraic computation. The same quantity can be derived using polar coordinates. We start with an arbitrary complex number $Z = a + bi = r\cos(\theta) + ir\sin(\theta)$, and square it to obtain

$$Z^2 = [r(\cos(\theta) + i\sin(\theta))]^2 = \left[re^{i\theta}\right]^2 = r^2e^{i2\theta} = r^2(\cos(2\theta) + i\sin(2\theta)).$$

Because of the use of polar coordinates, the relation between $Z = (r, \theta)$ and $Z^2 = (r^2, 2\theta)$ is now clear. The radius r of Z is squared, and its angle θ is doubled (Figure 2.39a). Thus, to get from a complex number to its square root we need to halve its angle and compute the $\sqrt{r}$.

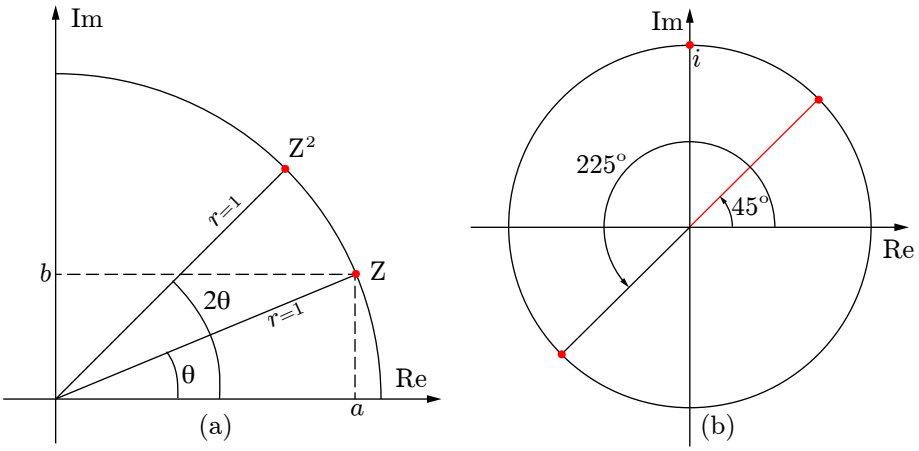


Figure 2.39: Square Root of Z.

This is for an arbitrary Z, but we are interested in the particular Z that satisfies $Z^2 = i$. The polar coordinates of i are $(1, 90°)$, so those of $\sqrt{i}$ should be $(1, 45°)$. When converted back to Cartesian coordinates, these result in

$$\sqrt{i} = \cos(45°) + i\sin(45°) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}i.$$

Equation (2.20) promises two square roots, and the other one can be obtained in polar coordinates by adding 180° to the 45° (Figure 2.39b). This results in

$$\cos(225°) + i\sin(225°) = -\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}i.$$

◇ **Exercise 2.43:** What are the $\sqrt{2i}$ and the $\sqrt{2+i}$?

◇ **Exercise 2.44:** (Funny.) Consider the following long, but elementary derivation:

$$\frac{1}{\sqrt{i}} \overset{(1)}{=} \frac{\sqrt{1}}{\sqrt{i}} \overset{(2)}{=} \sqrt{\frac{1}{i}} \overset{(3)}{=} \sqrt{\frac{1}{\sqrt{-1}}} \overset{(4)}{=} \sqrt{\frac{\sqrt{1}}{\sqrt{-1}}} \overset{(5)}{=} \sqrt{\sqrt{\frac{1}{-1}}} \overset{(6)}{=} \sqrt{\sqrt{-1}} \overset{(7)}{=} \sqrt{i}.$$

It suggests that $i = 1$. Where is the error?

• The next candidate to explore at this point is $\sqrt[i]{i}$. From the Argand diagram it is obvious that the polar coordinates of -1 are $(-1, \pi)$, which is why -1 equals $e^{i\pi}$. From this and from $\sqrt{-1} = i$ we deduce that

$$(-1)^{\frac{1}{i}} = \sqrt[i]{i} = \left[e^{i\pi}\right]^{\frac{1}{i}} = e^{\pi} \approx 23.1407.$$

Another complex computation that results in a real quantity.

• What is i^i? Applying the identity $a = e^{\ln a}$, we start with $i^i = \left[e^{\ln(i)}\right]^i = e^{i\cdot\ln(i)}$. We now recall that a complex number with polar coordinates (r, θ) can be written as $Z = re^{i\theta}$, which implies

$$\ln(Z) = \ln r + \ln[e^{i\theta}] = \ln r + i\,\theta.$$

However, in the Argand diagram, i is the point with coordinates $(0, 1)$, which translates to polar coordinates $(r, \theta) = (1, \pi/2)$. Based on this, we write $\ln i = \ln(1) + i\frac{\pi}{2} = i\frac{\pi}{2}$, which gives the final result

$$i^i = e^{i\cdot\ln(i)} = e^{i\cdot i\frac{\pi}{2}} = e^{-\frac{\pi}{2}} \approx 0.20788.$$

This is a real surprise; a complex power of a complex number turns out to be real. Incidentally, it turns out to be transcendental.

◇ **Exercise 2.45:** Based on the i^i example above, show how easy it is to compute the tetration i^{i^i}.

◇ **Exercise 2.46:** What is $1/i$?

◇ **Exercise 2.47:** Compute the value of $\sqrt{i} + \sqrt{-i}$.

• What is $\sin(i)$? This simple derivation starts with the two identities $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ and $e^{-i\theta} = \cos(\theta) - i\sin(\theta)$, which are subtracted to yield $e^{i\theta} - e^{-i\theta} = 2i\sin(\theta)$ or $\sin(\theta) = (e^{i\theta} - e^{-i\theta})/2i$. Selecting $\theta = i$ produces the final, complex result

$$\sin(i) = \frac{e^{i\cdot i} - e^{-i\cdot i}}{2i} = -\frac{1}{2}(e^{-1} - e)i = \left(\frac{e - e^{-1}}{2}\right) i = \sinh(1)i \approx 1.1752i.$$

A digression. We are familiar with the ordinary trigonometric functions sin and cos. They are defined in relation to circles, because the curve $(\cos t, \sin t)$ is the circle $x^2+y^2 = 1$. Similarly, the hyperbolic functions sinh and cosh (pronounced "shine" and "kosh",

respectively) are defined in relation to the hyperbola, and the curve $(\cosh t, \sinh t)$ is the right half of the unit hyperbola $x^2 - y^2 = 1$. The definitions are

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \text{and} \quad \cosh x = \frac{e^x + e^{-x}}{2}.$$

There are also hyperbolic tangent and cotangent. The argument of the hyperbolic functions can also be complex. (End of digression.)

- What is $\ln(i)$? We already know that i can be written in polar coordinates as $e^{i\frac{\pi}{2}}$, which implies that $\ln(i) = \ln\left(e^{i\frac{\pi}{2}}\right) = i\frac{\pi}{2}$. Simple!

◇ **Exercise 2.48:** What is $i!$?

◇ **Exercise 2.49:** Solve the equation $(-5)^x = 5$ in two ways.

- Now for something strange. We know that the first four powers of i (corresponding to exponents 1, 2, 3, and 4) are i, -1, $-i$, and 1, and that this repeats for the next groups of four consecutive exponents. However, this is true only for integer exponents. With a real exponent x, The power i^x can take other values, and here we will look at $i^x = 2$. What real power of i results in the integer 2? The simple answer is easily obtained when we consider the natural log of both sides of this equation

$$i^x = 2 \Rightarrow x\ln(i) = \ln(2) \Rightarrow x = \frac{\ln(2)}{\ln(i)} = \frac{\ln(2)}{i\pi/2} = -\frac{2i\ln(2)}{\pi}.$$

A simple check with mathematical software verifies that yes, $i^{-\frac{2i\ln(2)}{\pi}}$ equals 2.

The equation $x^i = 2$ is a related computation with i. The first step is $\sqrt[i]{x^i} = \sqrt[i]{2}$, which can be massaged to yield

$$x = 2^{\frac{1}{i}} = 2^{\frac{i}{i\cdot i}} = 2^{-i} = \left[e^{\ln 2}\right]^{-i} = e^{-i\ln 2} \overset{\text{Euler}}{=} \cos(\ln 2) - i\sin(\ln 2).$$

- An exercise in infinity. What is the sum Z of the infinite expression

$$Z = \sqrt{i + \sqrt{i + \sqrt{i + \cdots}}}?$$

Like many other problems involving infinite series and continued fractions, this example is also solved in a concise, elegant way, similar to the problem of Page 133, by noticing that any infinite part of this sum equals the entire sum Z, or that $Z = i + Z$. This produces the quadratic equation $Z^2 - Z - i = 0$ whose solution is $\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4i}$. We can stop at this point, but leaving i under the square root is inelegant and can be

avoided, so we continue and employ Equation (2.12) to write

$$\begin{aligned}\frac{1}{2}+\frac{1}{2}\sqrt{1+4i} &= \frac{1}{2}+\frac{1}{2}\left[\sqrt{\frac{1+\sqrt{17}}{2}}+\sqrt{\frac{-1+\sqrt{17}}{2}}i\right]\\ &= \frac{1}{2}+\left[\sqrt{\frac{2+2\sqrt{17}}{4}}+\sqrt{\frac{-2+2\sqrt{17}}{4}}i\right]\\ &= 0.5+0.800243+0.624811i = 1.300243+0.624811i.\end{aligned}$$

◇ **Exercise 2.50:** (Easy.) Given a positive integer n, what is $i^{n+1}-i^n$?

Tetration of i. The term "exponentiation" is used for repeated multiplication. Thus, x^b implies $\underbrace{x\times x\times\cdots\times x}_{b}$. Similarly, the term "tetration" (or hyper-4) means repeated exponentiation. The common notation for tetration is the left-exponent ${}^{n}x$, which denotes the power tower

$$\underbrace{x^{x^{\cdot^{\cdot^{x}}}}}_{n}.$$

Exercise 2.45 mentions the tetration of i, an interesting topic which is discussed here in some detail. We denote the nth tetration of i by q_i, and we already know the first few values of q_i, which are

$$q_1=i,\quad q_2=i^i=e^{\pi/2}\quad q_3=i^{i^i}=\cos\left[\frac{\pi}{2}e^{-\frac{\pi}{2}}\right]+i\sin\left[\frac{\pi}{2}e^{-\frac{\pi}{2}}\right],$$

and in general $q_{n+1}=i^{q_n}$. In order to compute the limit of q_n, we assume that it exists, i.e., that the sequence $\{q_n\}$ converges to a finite, complex-valued number Q, which equals the infinite power tower

$$Q=i^{i^{i^{i^{\cdot^{\cdot^{\cdot}}}}}}.$$

Since the tower is infinite, we try to write this equation in the form $Q=i^Q$. From the identity $i=e^{\frac{\pi}{2}i}$ we get $Q=e^{\frac{\pi}{2}iQ}$, and because Q is assumed to be complex (as a special case it may be real), we write it in the form $Q=a+bi$ to obtain

$$a+bi=e^{\frac{\pi}{2}i(a+bi)}=\cos\left[\frac{\pi}{2}a\right]e^{-\frac{\pi}{2}b}+i\sin\left[\frac{\pi}{2}a\right]e^{-\frac{\pi}{2}b},$$

which produces the two equations

$$a=\cos\left[\frac{\pi}{2}a\right]e^{-\frac{\pi}{2}b},\qquad b=\sin\left[\frac{\pi}{2}a\right]e^{-\frac{\pi}{2}b}.$$

Applying iterative methods, these can be solved to yield

$$Q=a+bi\approx(0.438283+0.360592i).$$

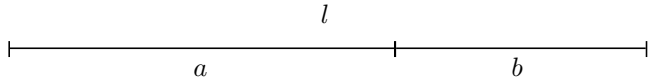
Figure 2.40 illustrates the result of computing the power tower tetration of i for 30 iterations. Part (a) of the figure displays the successive points in the complex plane, labeled and joined by edges. Part (b) shows just the points and it seems that they form three spirals. Part (c) proves that there really is just one spiral and the three spirals are only an optical illusion. The green circle is the vanishing point Q whose complex coordinates are $(0.438283, 0.360592)$.

Reference [redditI 20] includes a diagram similar to Figure 2.40a,b. `reddit` users `alexlozovsky` and `Chand_laBing` shed more light on the unexpected shape of the distribution of points in Figure 2.40. Here are some of their comments:

> Well, apparently, there is a reason why we can see 3, 5, 7, or 9 spirals. If we connect consecutive points on the plot for $n(i/k)$ and determine the angles between adjacent segments of the polygonal chain, we would find that the angles oscillate around some value $\alpha(k)$, which determines the "shape" of the polygonal chain.
>
> Thus, we get the following star-like polygrams:
> $\alpha(0.56) \approx 180/3 = 60°$ (which gives us triangular shape and 3 spirals),
> $\alpha(2.08) \approx 180/5 = 36°$ (which gives us pentagram and 5 spirals),
> $\alpha(4) \approx 180/7 = 25.7°$ (which gives us heptagram and 7 spirals),
> $\alpha(6.75) \approx 180/9 = 20°$ (which gives us enneagram and 9 spirals).
>
> So the answer to the question why does it form three distinct spirals? is because the angles between adjacent segments of the polygonal chain $\alpha(1) \approx 50.55°$, which is close enough to the triangular shape of the chain.
>
> You can play with it with the help of Wolfram Mathematica (or Wolfram Cloud) using the following function:
>
> ```
> T[z_]:=ComplexListPlot[NestList[N[z^#]&,0,1000],PlotRange->All]
> ```
>
> Executing `T[I/7]` would plot the graph for the base $i/7$.
>
> It can be proved that successive iterations are asymptotically rotated 129° about the sequence's limit. Since this is approximately 120°, every third iteration maintains approximately the same angle and lies on one of the three branches. The overshoot of 9° in part causes the spiraling.

2.11.4 Φ, The Golden Ratio

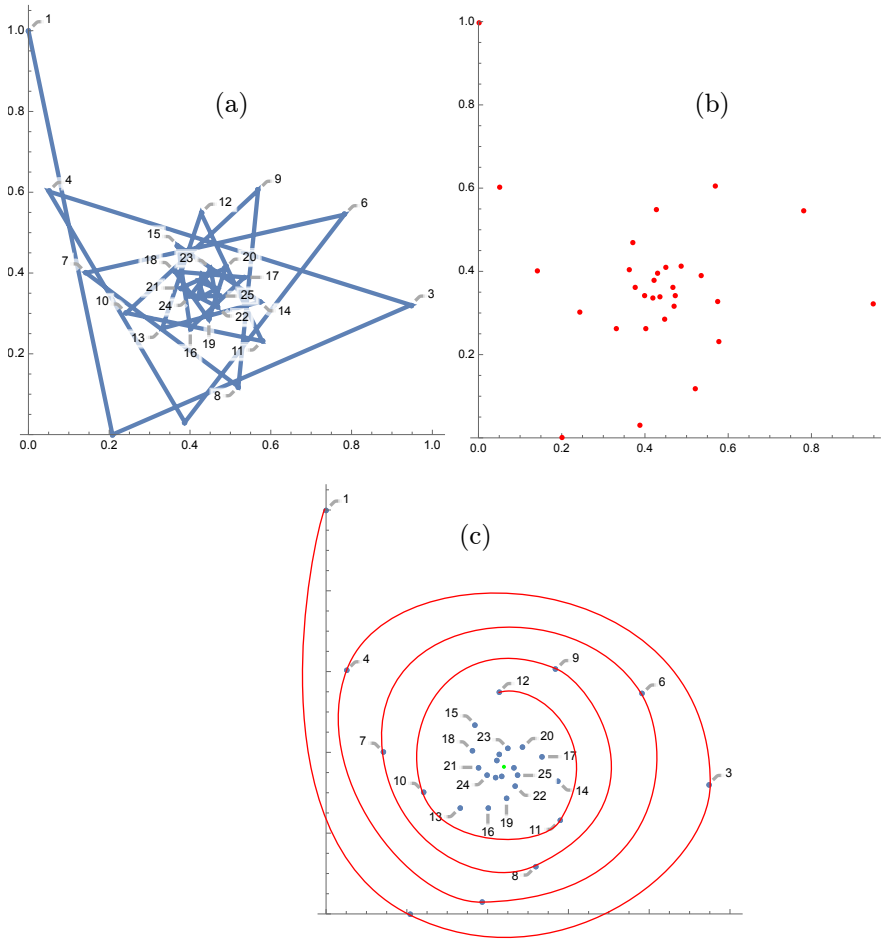
Start with a straight segment of length l and divide it into two parts a and b such that $a + b = l$ and $l/a = a/b$.



The ratio a/b is a constant called the Golden Ratio and is denoted by ϕ or φ (the Greek letter phi). It is one of the important mathematical constants, like π, i, and e, and was already known to the ancient Greeks. It is also known as the extreme and mean ratio and the divine proportion. Other names for it are golden mean, golden section, golden cut, golden proportion, golden number, medial section, and divine section. [WikiPhi 23] a general detailed reference to the golden ratio. See also Page 61.

Page 64 discusses the golden angle, a close relative of φ.

The simple construction above is only one of many ways to create this important number. Figure 2.41 illustrates three geometric constructions that contain ϕ and its




```
n = 30;
Tetration[a_, k_Integer] := Nest[Power[a, #] &, 1, k];
data = Table[N[Tetration[I, k]], {k, 1, n}];
ldata = Table[k, {k, 1, n}]; (*Label the points*)
ComplexListPlot[
 data -> ldata, Mesh -> Full, Joined -> True,
 PlotRange -> {Automatic, {0, 1}}]
ComplexListPlot[data, PlotRange -> {Automatic, {0, 1}},
 PlotStyle -> Red]

(*The following is proposed by reddit.com user alexlozovsky.
Try small values for k, such as  0.56, 0.58, 2.08, 4, and 6.75 to \
obtain different spirals.
T[z_]:=ComplexListPlot[NestList[N[z^#]&,0,100],PlotRange->All];
T[I/k]*)
```

Figure 2.40: The Tetration of i as a Spiral.

inverse. In part (a) of the figure, the parabola $y = x^2 - 1$ and the straight line $y = x$ intersect at (ϕ, ϕ) and $(1/\phi, 1/\phi)$. In parts (b) and (c), the ratio AC:BC equals ϕ. (Part (b) consists of a semicircle circumscribing a square. In part (c), the horizontal segment passes through the midpoints of the equilateral triangle's sides.)

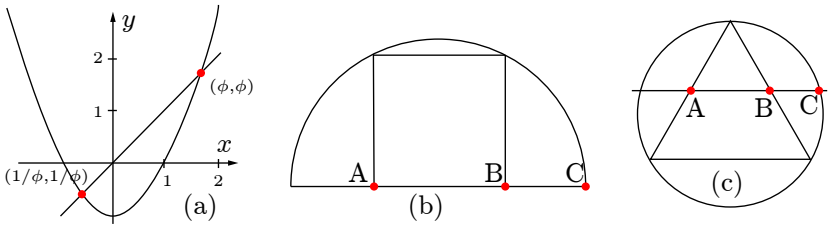


Figure 2.41: Three Geometric Constructions For ϕ.

It is often claimed that geometric figures can be made more pleasing to the eye if they involve this ratio, but such a claim to beauty is always in the eye of the beholder. One example of such a figure is the golden rectangle, whose sides are 1 and ϕ (or a and $a\phi$) long. Many classical buildings and paintings employ this ratio. [Huntley 70] is a lively introduction to the golden ratio. It illustrates properties such as

$$\phi = \sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}} \qquad \text{and} \qquad \phi = 1 + \frac{1}{1+\frac{1}{1+\frac{1}{\cdots}}}.$$

History

The golden ratio (ϕ or φ), its definition as a ratio, and its numeric value, were already known to the ancient Greeks. Even before Euclid, Pythagoras and his followers, while studying the properties of the pentagon around 540 B.C., came across the golden ratio which they considered special. This section is a short summary of the history of the terms "golden ratio," "golden mean," "golden number," and "golden proportion." Reference [Fowler 81] has more about this topic.

In Book VI, Proposition 30 of his great series the *Elements*, Euclid defines the golden ratio as follows "A straight line is said to have been cut in extreme and mean ratio when, as the whole line is to the greater segment, so is the greater to the less."

It seems that the term "golden ratio" first appeared in print in 1835 as the German *goldener schnitt*, in a book by Martin Ohm. It first appeared in English as "golden section" in the ninth edition of the *Encyclopaedia Britannica* (1875). Its first use in English as a mathematical term is found in the 1898 book *Introduction to Algebra* by George Chrystal.

In Italian, the term *divina proportione* is due to Luca Pacioli who chose it as the title of his book *La Divina Proportione*, (Venice, 1509). The Latin term *proportio divina* seems to have originated in a 1608 letter from Johannes Kepler who also used it later in

a 1611 book. The *Oxford English Dictionary* credits the term "golden mean" to D'Arcy W. Thompson, who used it on page 643 of his famous 1917 book *On Growth and Form*. The term "golden mean" is unfortunate, because it normally refers to the Aristotelian principle of moderation (to avoid extremes). The terms "golden rule," "golden number," and "golden proportion" are also used by various authors.

The following are a few of the many major references to the golden ratio, its history, its name, and its many properties:

- H. S. M. Coxeter, *Introduction to Geometry*, John Wiley, 1961. See Ch. 11.
- *The Fibonacci Quarterly*, published by The Fibonacci Association. See `fq.math.ca`
- Mario Livio, *The Golden Ratio: The Story of PHI, the World's Most Astonishing Number*, Crown, 2003.
- Gary B. Meisner, *The Golden Ratio: The Divine Beauty of Mathematics*.
- H. E. Huntley, *The Divine Proportion: A Study in Mathematical Beauty*, Dover, 1970. (Notice, the last equation on page 99 is wrong)
- Unfortunately, perhaps because of human nature, the golden ratio, which is "only" an important mathematical constant, has given rise to a vast body of literature that (often based on nothing more than wishfulness to see beauty where none exists) purports to find ϕ in many places in art, architecture, and nature. An excellent reference is [Livio 24]. This is why subsection 2.12.1 shows how the original definition of ϕ in terms of two equal ratios can be generalized to many extreme mean ratios whose geometric and algebraic properties resemble those of ϕ.

Claims similar to the following are not uncommon "We sometimes behave as if this ratio (ϕ) is nature's favorite number. It is divine and is blessed with an eighth wonder that makes it an ideal of beauty and perfection. Is it reasonable to claim that the ratios and proportions of successful works of art, design, and commerce are the chief secret behind their success? I believe that other features—such as composition, detail, and the correct use of colors—are more important. Golden ratio enthusiasts have tried to find their sweetheart in places such as the Mona Lisa, the Parthenon in Athens, and the logo of Apple computer. Even type design is sometimes purported to be beautiful if it includes this proportion. I say that ϕ is bullshit."

On the other hand, cooler heads point out that one definition of ϕ is as the limit of the ratio F_{n+1}/F_n of consecutive Fibonacci numbers. The larger n, the closer the ratio is to ϕ. Hence, there are infinite pairs of consecutive Fibonacci numbers that an artist, a designer, or an architect can use. None of them yields the exact value of ϕ, which is why measuring a work of art and finding its proportions very close to ϕ should not be a reason to assume that its creator had ϕ in mind.

◇ **Exercise 2.51:** Dear reader. The discussion on Page 90 shows that ϕ is the most irrational number. How special does this make ϕ?

The following quotation is from Clement Falbo, *The Golden Ratio: A Contrary Viewpoint*, reference [Falbo 05].

> Over the past five centuries, a great deal of nonsense has been written about the golden ratio, its geometry, and the Fibonacci sequence. Many authors

> make claims that these mathematical entities are ubiquitous in nature, art, architecture and anatomy. Gardner ... has shown that the admiration for this number seems to have been raised to cult status. Fortunately, however, there have been some recent papers, including Fischler ... in 1981, Markowsky ... in 1992, Steinbach... in 1997, and Fowler ... in 1982, that are beginning to set the record straight.

Finally, a long quotation from Mario Livio, *The Golden Ratio: The Story of Phi, the World's Most Astonishing Number*, (2003).

> Some of the greatest mathematical minds of all ages, from Pythagoras and Euclid in ancient Greece, through the medieval Italian mathematician Leonardo of Pisa and the Renaissance astronomer Johannes Kepler, to present-day scientific figures such as Oxford physicist Roger Penrose, have spent endless hours over this simple ratio and its properties. ... Biologists, artists, musicians, historians, architects, psychologists, and even mystics have pondered and debated the basis of its ubiquity and appeal. In fact, it is probably fair to say that the Golden Ratio has inspired thinkers of all disciplines like no other number in the history of mathematics.

The value of ϕ is easy to calculate. The basic ratio $l/a = a/b = \phi$ implies $(a+b)/a = a/b = \phi$, which, in turn, means $1 + b/a = \phi$ or

$$1/\phi = \phi - 1, \tag{2.21}$$

an equation that can also be written

$$\phi^2 = \phi + 1. \tag{2.22}$$

The latter equation is easy to solve, yielding $\phi = (1 + \sqrt{5})/2 \approx 1.618\ldots$. Both equations (2.21) and (2.22) have interesting properties that offer deeper insight into ϕ. The former shows that subtracting 1 from ϕ produces its inverse $1/\phi$, while the latter indicates that adding 1 to ϕ generates its square ϕ^2.

Being the root of an algebraic equation means that ϕ is an algebraic number, not transcendental. However, it is irrational, and is even the most irrational number, as shown on Page 90, in the sense of being as far as possible from any "nearby" rational number.

Equation (2.22) also generates an interesting sequence of consecutive powers of ϕ that relates to ϕ itself. From $\phi^2 = \phi + 1$ we derive

$$\begin{aligned}
\phi^3 &= \phi^2\phi = (\phi+1)\phi = \phi^2 + \phi = 2\phi + 1,\\
\phi^4 &= \phi^3\phi = (2\phi+1)\phi = 2\phi^2 + \phi + 1 = 3\phi + 2,\\
\phi^5 &= \phi^4\phi = (3\phi+2)\phi = 3\phi^2 + 2\phi = 5\phi + 3,\\
\phi^6 &= \phi^5\phi = (5\phi+3)\phi = 5\phi^2 + 3\phi = 8\phi + 5,\\
&\vdots\\
\phi^n &= F_n \cdot \phi + F_{n-1} = F_{n+1} + F_{n-1}.
\end{aligned} \tag{2.23}$$

Thus, the powers ϕ^n are related to products of the form $F_{n-1}\phi$; a strong relation to the Fibonacci numbers! The powers of ϕ can be said to contain "deep within them" the celebrated Fibonacci numbers. This should be compared with Equation (5.2), which relates the powers of ϕ to the Lucas numbers L_n.

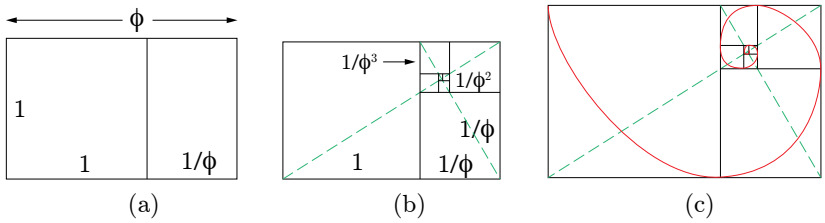


Figure 2.42: The Golden Ratio.

The equation $\phi = 1 + 1/\phi$ illustrates another unusual property of ϕ. Imagine the golden rectangle with sides $1 \times \phi$ (Figure 2.42a). Such a rectangle can be divided into a 1×1 square and a smaller golden rectangle of dimensions $1 \times 1/\phi$. The smaller rectangle can now be divided into a $1/\phi \times 1/\phi$ square and an even smaller golden rectangle (Figure 2.42b). When this process continues, the rectangles converge to a point. Figure 2.42c shows how the golden spiral—a special logarithmic spiral based on powers of ϕ, whose equation in polar coordinates is $r = \phi^{\frac{2\theta}{\pi}}$—can be drawn through corresponding sides of the rectangles.

The fact that the golden rectangle depends on ϕ may cause a confusion, because the aspect ratio of a rectangle may be the shorter side over the longer side, or the reverse. Thus, most people may agree that the aspect ratio of the golden rectangle is $\phi/1 = (\sqrt{5}+1)/2 \approx 1.618$, but some may insist that this ratio is $1/\phi = (\sqrt{5}-1)/2 \approx 0.618$. Even mathematics can sometimes be confusing.

◇ **Exercise 2.52:** Show how to draw a golden rectangle with a ruler and compass.

A geometric construction of ϕ. Figure 2.43 shows a simple, ruler and compass contraction of the irrational ϕ. Start with the green line of length $2L$. Create a vertical (red) copy of it, and draw the blue diagonal whose two parts are L and (the unknown) a.

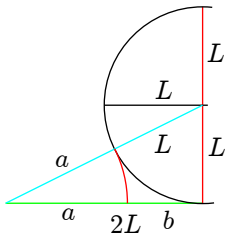


Figure 2.43: A Construction of ϕ.

In the resulting right triangle, the length of the diagonal is $\sqrt{(2L)^2+L^2}=\sqrt{5}L$, which is why $a=\sqrt{5}L-L$. Once we rotate a down to the green line, the ratio $2L/a$ can easily be computed:

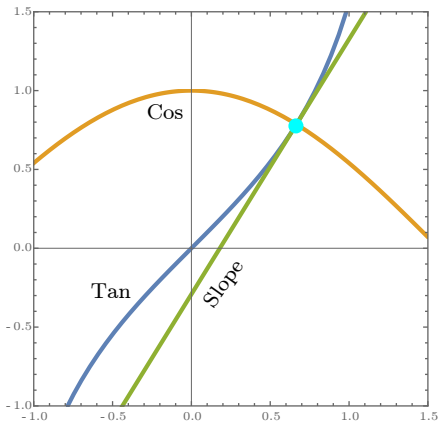
$$\frac{2L}{a}=\frac{2L}{L(\sqrt{5}-1)}=\frac{2}{(\sqrt{5}-1)}\frac{(\sqrt{5}+1)}{(\sqrt{5}+1)}=\frac{\sqrt{5}+1}{2}=\phi.$$

◇ **Exercise 2.53:** (Tricky, but mathematically simple.) Why does the golden ratio exist?

2.11.5 Unusual Appearances of Φ

Much has been written about the golden ratio and its close relatives, the Fibonacci numbers. They seem to appear everywhere, in nature, art, literature, and even in music (reference [FibonacciMusic 22]). This Subsection describes two little-known and unusual cases where this ratio rears its beautiful head.

■ The slope of $\tan x$. One reference for this is [DrBarker 24]. At the point the graphs of $\tan x$ and $\cos x$ intersect, the slope of $\tan x$ is ϕ. The coordinates of the point are $(0.666239, 0.786151)$ and the graphs are shown in Figure 2.44. The proof is straightforward.




```
f[x_] := 0.08x^4+0.25x^3-1.2x^2-1.6x+1.2;
Plot[{Tan[x], Cos[x], 1.618 x - 0.291824}, {x, -1, 1.5},
 PlotRange -> {{-1, 1.5}, {-1, 1.5}}, Frame -> True,
 AspectRatio -> 1]
```

Figure 2.44: Tangent and Cosine Meet.

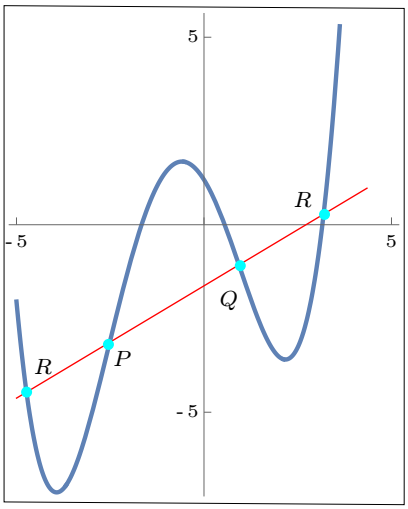
$$\text{At the intersection, } \cos\theta = \tan\theta = \frac{\sin\theta}{\cos\theta},$$
$$\cos^2\theta = \sin\theta,$$
$$\sin^2\theta + \sin\theta - 1 = 0,$$
$$4\sin^2\theta + 4\sin\theta = 4,$$
$$(2\sin\theta + 1)^2 = 5,$$
$$\text{So at the point, } \sin\theta = \frac{-1 \pm \sqrt{5}}{2},$$
$$\frac{d\,\tan\theta}{dx} = \frac{1}{\cos^2\theta},$$
$$\cos^2\theta = 1 - \sin^2\theta = \sin\theta, \text{ (since } \sin^2 x + \sin x - 1 = 0),$$
$$\frac{1}{\cos^2\theta} = \frac{2}{-1+\sqrt{5}} = \frac{2(\sqrt{5}+1)}{4} = \frac{\sqrt{5}+1}{2} = \phi.$$

■ The inflection points of a quartic polynomial. A quartic (degree-4) polynomial is an elementary function of the form $f(x) = Ax^4 + Bx^3 + Cx^2 + Dx + E$. Because of the two terms x^3 and x^4, the graph of this polynomial can take many different and complex shapes and can also have two inflection points.

An inflection point of a curve is a point where the curve changes its concavity. Imagine a point sliding down a curve in the form a large letter "S." While sliding in the top part of the curve, the point moves along a semicircle in a counterclockwise direction. After reaching the midpoint of the curve, the direction of the point changes to clockwise. The midpoint of this curve is an inflection point, a point where the curvature of the curve changes sign between positive and negative. At the inflection point, the curvature of the curve is zero; on one side of the point the curvature is positive, while on the other side the curvature is negative.

Points P and Q of Figure 2.45 are inflection points of the quartic polynomial $0.08x^4 + 0.25x^3 - 1.2x^2 - 1.6x + 1.2$, and the interesting feature that is discussed here has to do with the two points R. Those are the points on the curve that lie on the straight line that passes through P and Q. As long as the curve has two inflection points P and Q, the ratio PR/PQ always equals ϕ for both points R, regardless of the individual values of the five polynomial parameters and of the particular shape of the curve. Reference [GoldenQuartics 20], by Adi Mittal, presents an elegant proof of this little known fact.

Here is some more information about inflection points, for the benefit of inquisitive readers. The definition of an inflection point implies that this is a point where the curve moves in a straight line (a straight line that has shrunk to a single point). The slope of a curve $f(x)$ is its first derivative $f'(x)$, which can also be visualized as a curve. At an inflection point, where $f'(x)$ is straight, its slope, which is $f''(x)$, is zero, Therefore, given a curve $f(x)$, its inflection points can be located by computing its second derivative.




```
f[x_] := 0.08x^4+0.25x^3-1.2x^2-1.6x+1.2;
Plot[f[x], {x, -5, 5}, AspectRatio -> Automatic,
 Ticks -> {{-5, 0, 5}, {-5, 5}}]
```

Figure 2.45: A Quartic Curve with Three Points.

In the case of a quartic polynomial, the derivatives are

$$f(x) = Ax^4 + Bx^3 + Cx^2 + Dx + E$$
$$f'(x) = 4Ax^3 + 3Bx^2 + 2Cx + D$$
$$f''(x) = 12Ax^2 + 6Bx + 2C.$$

The second derivative is therefore a quadratic (degree-2) polynomial, so equating it to zero produces a quadratic equation whose solutions are

$$P = \frac{-3B - \sqrt{9B^2 - 24AC}}{12A}, \qquad Q = \frac{-3B + \sqrt{9B^2 - 24AC}}{12A}.$$

Once this point in the derivation in reached, the analysis is trivial. If the term $9B^2 - 24AC$ is negative, the curve has no inflection points. If this term is zero, there is one infection point. Otherwise, there are two such points.

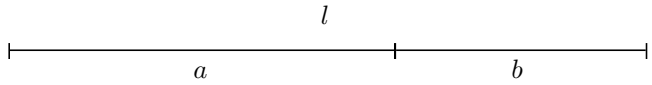
The following is an unexpected property of ϕ. The definite integral

```
Integrate[1./(1 + x^GoldenRatio)^GoldenRatio, {x, 0, \[Infinity]}]
```

evaluates to 1.

2.12 Complex Golden Ratios

The golden ratio ϕ is defined in Section 2.11.4 as the ratio a/b that is obtained when we start with a straight segment of length l and partition it into two parts a and b such that $a+b=l$ and $l/a=a/b$.



This section extends this definition by partitioning segment l in four different ways to obtain four ratios as follows

$$\phi=\frac{a}{b}=\frac{a+b}{a},\quad \phi_j=\frac{a}{b}=-\frac{a+b}{a},$$
$$\phi_i=\frac{a}{b}=\frac{a-b}{a},\quad \phi_k=\frac{a}{b}=-\frac{a-b}{a}.$$

Equation (2.22) serves as a guide to produce a set of four quadratic equations

$$\phi^2-\phi-1=0,\quad \phi_j^2+\phi_j+1=0,$$
$$\phi_i^2-\phi_i+1=0,\quad \phi_k^2+\phi_k-1=0.$$

These are easy to solve and produce the set of four general golden ratios, three of which are complex numbers

$$\phi=\frac{1+\sqrt{5}}{2},\quad \phi_j=\frac{-1+i\sqrt{3}}{2}=e^{2\pi i/3},$$
$$\phi_i=\frac{1+i\sqrt{3}}{2}=e^{i\pi/3},\quad \phi_k=\frac{-1-i\sqrt{3}}{2}=e^{-2\pi i/3}.$$

Each of the four golden ratios has a simple relation with its inverse as follows

$$\phi=1+\frac{1}{\phi},\quad \phi_j=-1-\frac{1}{\phi_j}\quad \phi_i=1-\frac{1}{\phi_i}\quad \phi_k=-1+\frac{1}{\phi_k}.$$

The recursive relation between the original golden ratio and the Fibonacci numbers has similar relatives which apply to all four general golden ratios

$$F_n=F_{n-1}+F_{n-2},\quad F_{j,n}=-F_{j,n-1}-F_{j,n-2},$$
$$F_{i,n}=F_{i,n-1}-F_{i,n-2},\quad F_{k,n}=-F_{k,n-1}+F_{k,n-2}.$$

The Binet formula, due to Jacques Philippe Marie Binet, provides a closed expression for the nth Fibonacci number in terms of ϕ and its inverse $\psi=-1/\phi$

$$F_n=\frac{\phi^n-\psi^n}{\phi-\psi}.$$

This formula applies to all four generalized golden ratios, where each has its own ψ inverse.

Section 2.11.4 also shows how ϕ can be expanded as an infinite nested square roots and an infinite continued fraction

$$\phi=\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}}\qquad\text{and}\qquad\phi=1+\frac{1}{1+\frac{1}{1+\frac{1}{\cdots}}}. \tag{2.24}$$

This also applies to the three complex golden ratios. All four can be expanded into continued fractions whose compressed definitions are

$$\begin{aligned}\phi&=[1:1,1,1\ldots],\quad \phi_j=[-1:-1,-1,-1,\ldots],\\ \phi_i&=[1:-1,-1,-1,\ldots],\quad \phi_k=[-1:1,1,1,\ldots].\end{aligned}$$

Finally, the four ratios can be expressed as nested square roots in the forms

$$\begin{aligned}\phi&=\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}},\quad \phi_j=\sqrt{1+i\sqrt{1+i\sqrt{1+i\sqrt{1+\cdots}}}},\\ \phi_i&=\sqrt{-1+\sqrt{-1+\sqrt{-1+\sqrt{-1+\cdots}}}},\quad \phi_k=\sqrt{-1+i\sqrt{1-i\sqrt{1-i\sqrt{1-\cdots}}}}.\end{aligned}$$

References for this section are [mathforum 11], [stackexchange1851698 23], and [stackexchange4856811 24].

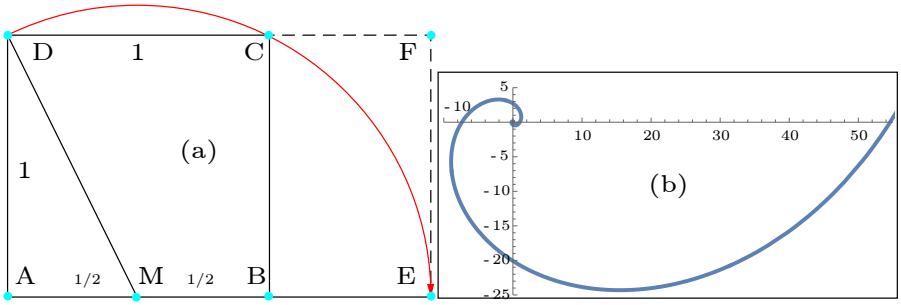
◇ **Exercise 2.54:** How can we construct imaginary and complex numbers that will be non-real "equivalents" of the golden ratio?

2.12.1 Generalizing the Golden Ratio

The original definition of ϕ, the golden ratio, is by means of ratios. This definition is generalized here using areas instead of ratios. Euclid himself, in Book III, Proposition 11 of his *Elements* explains how to construct ϕ geometrically. One only has to cut a given straight line so that the rectangle contained by the whole and one of the segments equals the square on the remaining segment. Figure 2.46a illustrates what this means (see also Exercise 2.52). Start with a 1×1 square $ABCD$ and place point M midway on segment AB. Rotate segment MD to determine point E.

From the right triangle ADM it is clear that $(DM)^2=1^2+(1/2)^2$, so $DM=\sqrt{5}/2$, but from the construction it is also obvious that $DM=ME$. Therefore $BE=ME-MB=(\sqrt{5}-1)/2$, which implies that $AE=AB+BE=1+BE=(\sqrt{5}+1)/2=\phi$. Our rectangle is golden, and satisfies the defining golden ratio $AE/AB=AB/BE$, which is similar to the relation $l/a=a/b$ (see the beginning of Section 2.11.4).

A well-known property of the golden ratio, illustrated in Figure 2.42b and in many other places, is that it can be partitioned into smaller and smaller subrectangles, all featuring the same ratio and forming a spiral. However—and this is the first "nail in the coffin of ϕ"—this property of the golden ratio is shared by almost all other rectangles, as the following statement makes clear:




```
PolarPlot[E^(2t/Pi), {t,-4Pi,5Pi},PlotRange->{{-10,55},{-25,5}}]
```

Figure 2.46: Golden Rectangle and Spiral.

Any rectangle, except the square, can be partitioned into subractangles that share the same ratio of length to height.

This fact is illustrated, following the notation of [Fowler 81], by the construction of Figure 2.47a, which is a generalization of Figure 2.46a. Start by selecting a positive integer n (3 in the figure). Construct a 1×1 square $ABCD$ and set point M_n at a distance $n/2$ from A. Construct segment DM_n and rotate it to determine point E. The extended segment AE has length $n+\sqrt{n^2+4}/2$, a quantity which was coined *noem* (nth order extreme mean) by Fowler. The generalized defining ratios (noems) are

$$\frac{M_nE+AM_n}{AB}=\frac{AB}{M_nE-AM_n},\quad \text{or}\quad \frac{AE}{AB}=\frac{AB}{GE},$$

which for $n=1$ equals the original golden ratio ϕ, a ratio which can now be termed the first order extreme mean. For $n=3$, Figure 2.47a includes point G such that $AM_n=M_nG$ as implied by the red arc. Since AM_3 has length $3/2$, we conclude that the length of noem AE, the third order extreme mean, is

$$\begin{aligned}GE&=M_3E-M_3G=M_3D-AM_3\\&=\sqrt{1^2+1.5^2}-3/2=\sqrt{13/4}-3/2=(\sqrt{13}-3)/2,\\AE&=AM_3+M_3G+GE=3/2+3/2+\sqrt{13}/2-3/2=(3+\sqrt{13})/2.\end{aligned}$$

◇ **Exercise 2.55:** What is the noem for $n=2$?

The generalized golden ratios (or extreme means) are not limited to integers, and can be obtained, following [Fowler 81], from Figure 2.47b for any real number p. The result is denoted by r and is dubbed *poem* (pth order extreme mean). It starts with

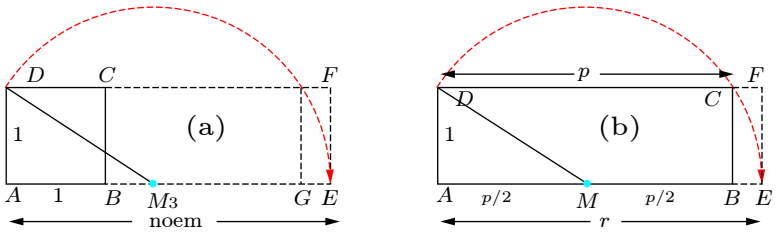


Figure 2.47: Noem and Poem Rectangles.

a $p\times 1$ rectangle $ABCD$, where side AB has length p. Point M, the midpoint of AB, is constructed, and segment MD, whose length is $\sqrt{p^2+4}/2$, is rotated to determine point E. The length of AE is $r=(\sqrt{p^2+4}+p)/2$, which yields the length of BE as $BE=AE-AB=(-p+\sqrt{p^2+4})/2$. The latter value equals $1/r$ because

$$\frac{\sqrt{p^2+4}+p}{2}\times\frac{-p+\sqrt{p^2+4}}{2}=\frac{(p^2+4)-p^2}{4}=1,$$

so we conclude that $r-p=1/r$ or $r^2=pr+1$, which is the generalization of the basic golden equations (2.21) and (2.22). Similarly, the relation $AE/EF=FE/BE$ is the generalization of the defining golden relation $l/a=a/b$, which appears at the beginning of Section 2.11.4.

A generalized golden rectangle for $r=1.4$ is shown in Figure 2.48, making it clear that this rectangle is not much different from the one for ϕ, shown in Figure 2.42. To construct this rectangle, start with a 1×1 square (small brown) and append a 0.4×1 rectangle to it (white). This is already a generalized golden rectangle for $r=1.4$, and it can be extended indefinitely by appending bigger and bigger squares (pink, brown, orange, cyan, blue, green, and red in the figure) in a spiral pattern.

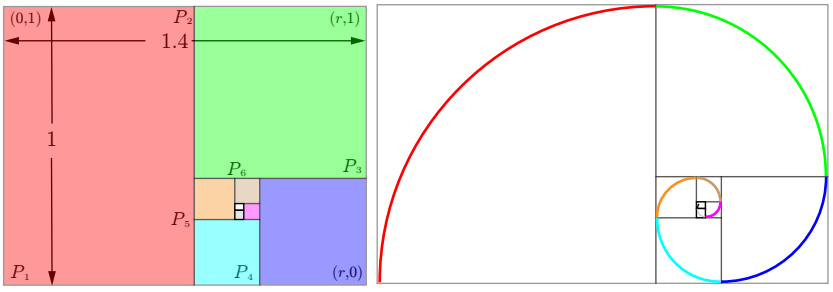


Figure 2.48: A Golden Rectangle with Ratio 1:1.4.

Reference [Falbo 05] derives the coordinates of various strategic points on such a

rectangle, and provides the list

$$\begin{aligned}
P_2 &= (r - 1/r, 1),\\
P_3 &= (r, 1/r^2),\\
P_4 &= (r - 1/r + 1/r^3, 0),\\
P_5 &= (r - 1/r, 1/r^2 - 1/r^4),\\
P_6 &= (r - 1/r + 1/r^3 - 1/r^5, 1/r^2),
\end{aligned}$$

which converges at

$$\left(\frac{r^3}{1+r^2}, \frac{1}{1+r^2}\right).$$

There are other constructions that lead to rectangles whose ratios converge to ϕ. A simple example is to start with a rectangle of sides $a_0 > b_0$ and apply the simple recursion $a_{n+1} = b_n$ and $b_{n+1} = a_n - b_n$ to construct smaller rectangles. The ratio r_n of sides of consecutive rectangles is derived by

$$\frac{1}{r_{n+1}} = \frac{b_{n+1}}{a_{n+1}} = \frac{a_n - b_n}{b_n} = \frac{a_n}{b_n} - 1 = r_n - 1,$$

and this relation approaches the limit $r^2 - r - 1 = 0$ whose familiar solution is $r = (1 + \sqrt{5})/2 = \phi$, thereby proving that this construction extends the golden rectangle.

A similar example is to start in the same way, and apply the simple recursion $a_{n+1} = a_n+, b_n$ and $b_{n+1} = a_n$, to construct bigger rectangles with ratios $r_{n+1} = 1 + 1/r_n$, that also converge to ϕ.

One can therefore claim that there is nothing special about the golden rectangle, but there is still something special about ϕ, because it is the limit of the ratios of so many general rectangles.

Notice that the golden rectangle, as well as the rectangles described earlier, are self-similar, which is why we can claim that ϕ is somehow associated with self similarity.

Figure 2.48 also illustrates the generalized golden spiral that fits inside this rectangle. In the figure, it is easy to see how this spiral was created out of several circular arcs. The golden spiral, original or generalized, is self similar, because the golden rectangle in which it fits can be extended indefinitely by appending more squares.

The golden spirals are special cases of logarithmic (equiangular) spirals, a family of curves that is described by the general polar equation $r = ae^{b\theta}$, or equivalently,

$$\theta = \frac{1}{b}\ln(r/a),$$

where a is the radius for $\theta = 0$, and b determines the shape of the curve. The golden spiral is obtained when the ratio r/a equals ϕ. Also, b has to satisfy $e^{b\theta_r} = \phi$, where θ_r is a right angle (either 90° or $\pi/2$). In this case, the expression $e^{b\theta_r} = \phi$ becomes

$$b = \frac{\ln\phi}{\theta_r} = \begin{cases} 0.0053468 & \text{for degrees} \\ 0.3063489 & \text{for radians.} \end{cases}$$

The original (not generalized) golden spiral is depicted in Figure 2.47b.

An Extended Fibonacci sequence

The Fibonacci sequence, Section 5.6, is based on the relation $F_n = F_{n-1} + F_{n-2}$, which is extended here to become

$$x_n = px_{n-1} + x_{n-2}, \tag{2.25}$$

with $x_0 = x_1 = 1$, $n \geq 2$, and for p that satisfies $p = r - 1/r$ for a positive real r. Reference [Falbo 05] lists closed expressions for the general term of this sequence, but even following the recursive relation (2.25), it is easy to see that the first few terms are

$$x_2 = p + 1, \quad x_3 = p^2 + p + 1, \quad \text{and} \quad x_4 = p^3 + p^2 + 2p + 1.$$

For $p = 1$, this is the original Fibonacci sequence, but for other values of p, the resulting sequence features properties that permit us to consider it a valid extension of the basic sequence.

We start with the basic equation of the golden ratio $x^2 - x - 1 = 0$ and extend it, in the spirit of relation (2.25), to $x^2 - px - 1 = 0$, a quadratic equation whose solutions (because $p = r - 1/r$) are r and $-1/r$, as can be seen by direct substitution. The following properties derive from relation (2.25):

1. Square of the nth term. It is well known that the original Fibonacci sequence has the unusual property that the square of any term is always one less or one more than the product of its immediate two neighbors. Thus, for example, $3^2 = 2 \cdot 5 - 1$, $5^2 = 3 \cdot 8 + 1$, and $8^2 = 5 \cdot 13 - 1$. This property is reflected in relation (2.25), because it can be proved by induction that for any three consecutive terms of it, the square of the middle term is always p more or p less than the product of its successor and predecessor.

2. Square roots and continued fractions. Equation (2.24), duplicated here, illustrates how ϕ can be expanded as an infinite nested square roots and an infinite continued fraction

$$\phi = \sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}}} \qquad \text{and} \qquad \phi = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{\cdots}}}. \tag{2.24}$$

Relation (2.25) yields the corresponding extensions

$$\phi = \sqrt{1 + p\sqrt{1 + p\sqrt{1 + p\sqrt{1 + \cdots}}}} \qquad \text{and} \qquad r = p + \frac{1}{p + \frac{1}{p + \frac{1}{\cdots}}}.$$

In summary, the above generalizations of the golden ratio, the golden rectangle, and the golden spiral seem to somehow reduce the importance of the golden ratio and demote it somewhat from its traditional prominent position. Those who see ϕ in art, architecture, and in nature and commercial design, should at least consider some of the extensions proposed here. There may, perhaps, be some application where an extension of ϕ may serve better than the original ϕ.

Poem on the Golden Ratio

In ancient times, a ratio was found,
A hidden thread through forms unbound,
From nature's hand and mind's keen eye,
Phi's beauty takes to the sky.

Golden mean, divine proportion,
In flowers' bloom, in ocean's motion,
From galaxies to nautilus shell,
In art and life, it casts its spell.

A line divided, so precise,
A harmony in every slice,
The larger part to whole relates,
A magic math that fascinates.

Fibonacci's sequence, simple and pure,
Unfolds a pattern, allure so sure,
Add two terms, the next reveals,
A spiral's grace, its charm appeals.

In Parthenon and pyramids grand,
Phi's presence lies, like nature's hand,
Da Vinci's art, a master's stroke,
In golden lines, his secrets cloak.

A leaf's unfurl, a petal's grace,
The curve of space, in endless chase,
DNA's twist, life's very code,
In golden spirals, life's road.

In music's chord, in symphony,
A golden thread of harmony,
Architects and poets know,
In golden ratio, ideas flow.

Beyond mere numbers, a cosmic link,
In every form, a common ink,
Phi connects the great unknown,
In every seed its pattern sown.

A universe of stars and space,
In golden spirals, we find our place,
From smallest cell to grand expanse,
Phi's elegance, a timeless dance.

In every form and every fold,
A story of the world is told,
Through golden ratio, life aligns,
A sacred script in endless lines.

So trace the spiral, round and round,
In nature's code, Phi is found,
A wondrous math, forever bright,
In golden curves, the world's delight.

By `chatGPT 4o`, May 2024.

Why are numbers beautiful? It's like asking why is Beethoven's Ninth Symphony beautiful. If numbers aren't beautiful, nothing is.

—Paul Erdős.

2.13 Approximating Formulas

Irrational numbers cannot be fully written with just digits, so we use letters and other symbols such as e, π, and $\sqrt{\cdots}$. The best that can be done to represent an irrational number is to write many digits or to develop a compact expression that yields an excellent approximation. One such expression, a surprising approximation of Euler's number e, is explained here.

The following definition is from Wikipedia: A pandigital number is an integer that in a given base has, among its significant digits, each digit used in the base at least once. The binary number 10_2, for example, is pandigital, as are also 201_3, 20102_3, and 0123456789_{10}. A similar concept is a pandigital formula. This is a mathematical expression that uses each digit (of its base) at least once, together with arithmetic operations and perhaps also variables.

With this in mind, it may come as a surprise to learn that the following pandigital formula provides an insane approximation of the all-important Euler's number e

$$\left[1+9^{-4^{6\cdot 7}}\right]^{3^{2^{85}}}. \tag{2.26}$$

Computing this formula results in more than 18×10^{24} (18 septillion) correct first digits of e; astonishing!

This esoteric, pandigital formula, due to Richard Sabey in 2004, is not a wild guess but the result of careful thinking and experimenting. It starts with the observation that one of the definitions of e is the limit of the simple, unusual expression $(1+1/n)^n$. Sabey (likely after several tries) started with the simple power $N=3^{2^{85}}$ and manipulated it to become

$$N=3^{2^{85}}=9^{2^{84}}=9^{4^{42}}=9^{4^{6\cdot 7}}.$$

This basic application of the rules of exponentiation expresses the same N with (on the left) the digits 3, 2, 8, 5, and on the right with the digits 9, 4, 6, and 7. Only a 0 and a 1 are needed in order to arrive at pandigital perfection.

With this in mind, the expression $(1+1/N)^N$ can be written

$$\left[1+\frac{1}{N}\right]^N=\left[1+\frac{1}{9^{4^{6\cdot 7}}}\right]^{3^{2^{85}}},$$

which no longer seems mysterious and has lost the incomprehensibility of Equation (2.26).

There does not seem to be a direct reference to Richard Sabey, but his formula is discussed in several places on the Internet. It is possible to get better approximations to e by starting with even larger numbers such as $N=9^{9^{9^9}}$, but the particular choice of $3^{2^{85}}$ makes sense because it leads to a pandigital formula.

◇ **Exercise 2.56:** What happened to zero? The Sabey formula is not pandigital because it misses the digit 0.

Reference [kexue.fm 23] is a summary of this and many other formulas (not pandigital) that approximate several mathematical constants.

The following pandigital formula, due to B. Ziv, reference [Weisstein-Ziv 23], produces an approximation of π, but it is only accurate to 10 decimal places

$$2^{5^{0.4}} - 0.6 - \left[\frac{0.3^9}{7}\right]^{0.8^{0.1}}.$$

When computed at high precision, it yields $3.1415926535|904\ldots$, where only the first 10 digits are correct.

Another pandigital formula, due to G. W. Barbosa, produces a 17-digit approximation of π

$$\frac{\ln\left\{[2\times 5! + (8-1)!]^{\sqrt{9}} + 4! + (3!)!\right\}}{\sqrt{67}}.$$

One could claim that this is also weakly pandigital, because the $(8-1)$ is a trick employed to obtain the digit 8.

There are , of course, much more accurate approximations of π, but they are not pandigital.

2.14 Cyclic Numbers and Metadromes

We start with the inverses of the decimal digits

$$1/2 = 0.5,\ 1/3 = 0.333\ldots,\ 1/4 = 0.25,\ 1/5 = 0.2,\ 1/6 = 0.1666\ldots,$$
$$1/7 = 0.142857142857\ldots,\ 1/8 = 0.125,\ 1/9 = 0.111\ldots.$$

Four of these inverses are infinite, repeating fractions. Of those, 1/7 is different, because its repeating part is the integer 142857. It is this property of 1/7 that turns 142857 into a cyclic number. Here is what it means.

Let's first multiply 1/7 by 10. On one hand this equals $1.42857142857\ldots$ and on the other hand it equals 10/7, which is also $1 + 3/7$. Thus $3/7 = 0.42857142857\ldots$, a fraction whose six most-significant digits 428571 are a permutation of 142857. Similarly, multiplying 1/7 by 132 results in 18,857,124, which produces, after separation and addition, $18 + 857124 = 857142$, another permutation.

Two more examples are (1) $142857 \times 13 = 1857141$. Separating this product into $1 + 857141$ and adding produces 857142, another permutation. (2) $142857 \times 316 = 45142812$. Separating this product into 45+142812 and adding produces the permutation 142857.

However, when we try this with 84, which is a multiple of 7, the result is $142857 \times 84 = 11999988$, which, after separation and addition, yields $11 + 999988 = 999999$. In general, when multiplying 1/7 by a multiple $7a$, the result is a, which can be any number and has nothing to do with the unusual properties of the inverse 1/7.

The cyclic number 142857 is based on 7 which is prime. There are infinitely many prime numbers p that give rise to cyclic numbers. The inverse $1/p$ of such a prime has the form

$$\frac{1}{p} = 0.[\text{cycle}][\text{cycle}]\ldots,$$

with $p-1$ digits per cycle. Multiplying such an inverse by 10^{p-1} shifts it $p-1$ digits to the left, obtaining

$$\frac{10^{p-1}}{p} = [\text{cycle}].[\text{cycle}][\text{cycle}]\ldots.$$

When we subtract these two expressions, the result is

$$\frac{10^{p-1}}{p} - \frac{1}{p} = \frac{10^{p-1}-1}{p} = [\text{cycle}],$$

which is why each cycle is of the form $(10^{p-1}-1)/p$, and an integer of the form $10^{p-1}-1$ consists of $(p-1)$ 9's.

Metadromes. An integer is a metadrome in a given base b if its digits are in strictly increasing order in that base. For example, 1234, 268 and 13479 are all metadromes in base 10. A close cousin is the plaindrome (not to be confused with palindrome), an integer whose digits are in nondecreasing order, such as 2334779. Second cousins are integers whose digits are in strictly decreasing (or nonincreasing) order.

The total number metadromes in base b is 2^{b-1}. In base 2 there are two metadromes, namely 0 and 1 (because 01_2 equals 1). In base 3 there are $2^{3-1} = 4$ metadromes (0, 1, 2, and $12_3 = 5$). In base 10 there are $2^{10-1} = 512$ metadromes, from 0 to 123456789. Thus, the first few metadromes are 0, 1, 5, 27, 194, 1865,....

A Poem on Real Numbers

In realms of numbers, vast and wide,
Two worlds within the mind abide,
The rational, precise and clear,
The irrational, strange and near.

From counting sheep to measuring lands,
Rationals serve our clear demands,
A fraction's form, a decimal's grace,
A number line, they firmly place.

One-half, three-quarters, simple, neat,
In finite terms, they stay discreet,
With patterns that are easy to trace,
In rationals, we find our space.

But then we delve to deeper grounds,
Where numbers hum with endless sounds,
Irrationals, where logic bends,
A mystery that never ends.

Pi circles round in endless quest,
An unbound number, never at rest,
Its digits dance without repeat,
A constant curve, a rhythmic beat.

Square roots of primes, a wondrous sight,
Where rationals can't find their light,
In endless streams, they flow unbound,
In them, infinity is found.

The golden ratio, Phi's bright glow,
In nature's art, its numbers flow,
A spiral's curve, a petal's span,
Irrational, a cosmic plan.

Yet both these realms, in tandem, lie,
In every thought, in every sky,
They weave together, form and void,
In harmony, they are deployed.

From simple sums to vast expanse,
In rationals, we take our stance,
But in the infinite's grand domain,
Irrationals break the chain.

They whisper secrets of the vast,
Of time and space, of future, past,
In endless digits, they confide,
A universe that won't divide.

Thus numbers dance in endless play,
Both rational and wild array,
In every thought, in every line,
A balance struck, a truth divine.

So honor both, the clear and wild,
In numbers' realm, the free and styled,
For in their dance, our minds expand,
In rational and irrational, we understand.

By `chatGPT 4o`, May 2024.

[Numbers] are the only characters which can actually be "read" by all people of this earth, regardless of language.

—Hermann Zapf, *(1985)*.



3
Symmetry

Symmetry comes from the Latin words sym (meaning together) and metron (meaning measuring). This term is frequently utilized in daily life to evoke ideas of aesthetics, beauty, harmony, and balance. It pertains to our inclination to establish, or at least understand, organization and attractiveness, and maybe even flawlessness. Symmetry is a characteristic that an object can have in mathematics. It is a process that keeps the object unchanged, restoring it back to its original state, position, and alignment, despite the majority of its points shifting. A circle exhibits strong symmetry as it appears the same after any rotation or reflection. Similarly, a square can maintain its shape and orientation when rotated in increments of 90° or reflected across four different axes.

This leads to the realization that symmetry is somehow related to transformations. Imagine a magician holding an object. He asks you to close your eyes while he does something to the object. When you open your eyes, the object looks the same. Did he do nothing? Possibly, but he could also have transformed the object through one of its symmetry transformations, perhaps by rotating or reflecting it. Thus, in some sense, symmetry is a set of transformations, but in practice, the term symmetry also has an esthetic meaning that cannot be defined just by considering transformations. An interesting example is palindromes. Words such as `refer`, `rotator`, `deified`, and `racecar`, which are spelled the same left-to-right and right-to-left. When encountering such a word and realizing that it is a palindrome and therefore symmetric, we also feel something. A feeling that has to do with esthetics and cannot be defined rigorously. A similar example is the fact, discovered by researchers, that people and animals prefer those with a symmetric face. Beauty, as we know, is in the eye of the beholder, but a lopsided mouth or a broken nose are very noticeable. When someone declares "I love broken noses," people generally think, "a weirdo."

Symmetry also has practical applications. An engineer designing a bridge may decide to give it a four-fold symmetry, because he then has to do only a quarter of the computations needed. Crystals often have rich symmetries, which is then magnified in

a large piece of material made from the crystal. For all practical purposes, such a chunk of material may be considered infinite.

Symmetry is also related to art, because a great many art works are either symmetric or include elements of symmetry. It is this relation which raises the question of why living objects tend to exhibit certain types of symmetry, mostly bilateral. The best answer that we can offer is natural selection and how it works. It is generally believed that natural selection works blindly. When life starts somewhere, it starts with unicellular forms that slowly grow but remain essentially shapeless blobs. As those living beings become more complex and start moving on their own, be it in water, land, or in the air, they may benefit from certain symmetries, so those individuals that feature such symmetries survive and may become more complex.

One symmetry that is beneficial to movement is a front-back axis, where the front has sensory organs. The difference between top and bottom is also important for a moving animal. In a land animal, the top must be able to resist attack, while the bottom should contain the means of locomotion. On the other hand, there is nothing to prefer left over right or the opposite. This must be the reason why the interiors of current animals exhibit a weak left-right bilateral symmetry (our hearts are slightly to our left), while the exteriors benefit from strong left-right symmetry of the legs, fins, or wings.

Many animals and flowers (butterflies and orchids come to mind) have a bilateral symmetry; their left and right halves are identical or almost so. In contrast with the general concept of symmetry, which is vague and not rigorously defined, bilateral symmetry is geometric and precise. An object has bilateral symmetry with respect to a plane P, if it is identical to its reflection about P. Rotational symmetry is similarly also easily defined. An object has complete rotational symmetry around an axis a if any rotation about a carries the object into itself. Because of the paucity of such objects (only the circle and sphere), we also talk about incomplete rotational symmetry, the kind that transforms an object into itself only for certain rotation angles. A starfish immediately comes to mind.

◇ **Exercise 3.1:** What symmetry is featured in Figure 3.1?

At the opposite of symmetry there is asymmetry, which very often is part symmetry or broken symmetry rather than the complete lack of symmetry. We can say that if symmetry implies rest, then asymmetry implies motion; if symmetry implies law and binding, then asymmetry implies accident and loosening; and if symmetry implies rigor and cooperation, then asymmetry implies leniency and discord. Thus, (1) the two front towers of many Gothic cathedrals are different, thereby breaking the bilateral symmetry of the rest of the structure, and (2) the human face is partly symmetrical.

Symmetricity. Many English speakers claim that symmetricity is not a word, but some propose to validate its use in certain contexts, similar to words such as transitivity and reflexivity. Wiktionary mentions the term symmetricity and defines it as the condition of being symmetric, or the extent to which, or the manner in which something is symmetric.

Chirality. An object or a figure is chiral (has chirality) if it cannot be mapped to its mirror image by rotations (in the plane) and translations alone. Familiar examples are hands, shoes, clock faces, non-isosceles triangles, and screws or helices. Intuitively,

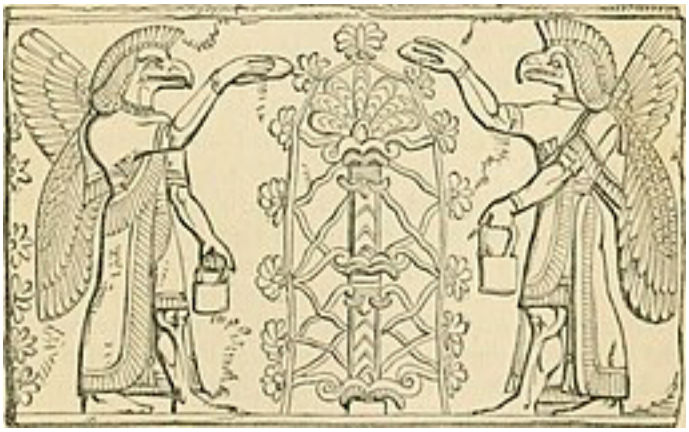


Figure 3.1: Symmetry in Ancient Art.

chirality is the lack of mirror symmetry, while achirality, its opposite, means the existence of mirror symmetry. A chiral object has no reflection symmetries. An object with such symmetries is achiral.

◇ **Exercise 3.2:** What are examples of left-handed threads?

Why are we attracted to symmetry? A psychologist might claim that symmetry represents order, and that we intuitively look for order in this unfamiliar and frightening universe we find ourselves in.

> By the word symmetry... one thinks of an external relationship between pleasing parts of a whole.
>
> —Johann Wolfgang von Goethe.

3.1 A bit of History

Before we delve into the details of symmetric objects and transformations, here is a short survey of a few important milestones in the development of the study of symmetry over the centuries. Perhaps the first person who noticed the beauty and usefulness of symmetry was the unknown inventor of the wheel. This familiar and practical object has infinite rotational symmetry. Even more, a two-dimensional wheel, a circle, which is infinitely thin and has no front and back, also has infinite reflection symmetry. It can be flipped about any axis that passes through its center.

The next major step in understanding symmetry was taken by the ancient Greeks, who discovered the five Platonic solids, those highly symmetric objects, the study of whose properties did much to advance Greek mathematics.

Much later in history, many important contributions to symmetry were made by the anonymous muslim artists who decorated the floors, walls, and ceilings of the Alhambra in Granada. Perhaps because of religious restrictions they never depicted images of

people or animals. All their creations are ornamental and many feature complex symmetries. In 1922, the Alhambra symmetries inspired Maurits Escher and became the source of his interest and future work on tessellations (regular divisions of the plane).

Skipping forward, we meet the young Évariste Galois, who as a teenager originated the branch of mathematical groups, the abstract objects used today to express, denote, and study symmetries.

But the real revolution in our understanding of symmetry came in the 20th century, where group theory became a powerful tool in the hands of many mathematicians. The study of symmetries and their properties and representations has culminated, in 1985, with the publication of the *ATLAS of Finite Groups*, a seminal work by several mathematicians, which lists all the finite simple groups, including much information about each.

Symmetry

Symmetry is a balance
Between two equal parts
It is a harmony
That pleases our eyes and hearts

Symmetry is a pattern
That repeats itself in space
It is a structure
That gives order and grace

Symmetry is a beauty
That we often admire
It is a reflection
That sparks our desire

Composed by ChatGPT Jan 2024

3.2 Symmetry Groups

In a mathematical sense, symmetry is a set of transformations. Several different objects may feature the same symmetries, which is why it is useful to develop a mathematical language or notation that captures the essence of symmetry by expressing symmetries in terms of transformations. This language has become known as groups or group theory. The collection of symmetries of any object is a group, and every group expresses the symmetries of some object. Thus, a mathematical group is a set with structure. It encapsulates algebraically the intuitive notion of the symmetry of an object, where by symmetry we mean an operation that leaves the object unchanged.

In other words, a symmetry group is a mathematical structure that characterizes the symmetries of an object, both 2D and 3D. It is the set of all the isometries that a leave the object unchanged. (Isometries are transformations that preserve the distance between any two points.)

Perhaps the first question that comes to mind, when we hear about symmetry groups is how to actually "see" such a group. Can we have a diagram, a list, a table,

or a mathematical expression that would visualize such a group. There are such tools, and some of them, such as the Cayley table (see Table 3.5) and the orbifold notation of Section 3.3 are explained in this book.

The advantage of applying mathematics to the study of symmetry is that objects may exist in higher dimensions, more than just two or three. Such objects cannot be visualized, but can still be investigated and their symmetries identified and summarized in a symmetry group. An extreme example of this power of groups is the so-called monster group, first discovered in 1973 and then constructed in 1980. This vast group expresses the symmetries of an object (a polytope) that "exists" in a 196,883-dimensional space.

A polygon is a plane (two-dimensional) convex shape bound by straight lines. A polyhedron (plural: polyhedra or polyhedrons; from Greek poly "many," and hedron "base or seat") is a three-dimensional shape with flat polygonal faces, straight edges and sharp corners or vertices. A polytope is the generalization of three-dimensional polyhedra to any number of dimensions.

In general, polygons may be considered two-dimensional polytopes, and polyhedra may be considered three-dimensional polytopes. Thus, the term polytope is most general and, in the spirit of Tolkien, we may refer to it as one term to rule them all. One of the simplest higher-dimensional polytopes is the tesseract, the four-dimensional cube.

So a group captures the essence of symmetry by expressing symmetries in terms of transformations, but what are these transformations? Let's start with scaling. Mathematically, an object can be stretched or shrunk or in general scaled by a certain factor. Being familiar with symmetry in real life, we intuitively feel that scaling is not a symmetry transformation, and a little thinking shows why. Scaling is not a symmetry because it changes the distance between points of the object. Scaling changes the size of the object but not its overall look, and yet we feel that this should not be considered a symmetric transformation. On the other hand, transformations such as rotation, reflection, and translation preserve distances, which is why they, or at least some of them, may be considered symmetry transformation.

This is why in mathematics, symmetry is an isometry (a function that preserves the distance between any two points) of the Euclidean space. Such a function may move points of the object, but leaves unchanged the footprint, the general shape, or the topology set of the points that constitute the object.

Symmetry is found in many mathematical objects, not just in geometric objects. Shifting every real number up by an amount a preserves the order of the numbers, as well as their relative distances, which is why this operation is a symmetry. Shifting the real numbers by $a = 1$ is a special case where integers remain integers. This is even more symmetric. Another example of symmetry in mathematics is swapping x and y in $x^3 + y^3 + z^3$. In music, we perform a symmetric transformation when transposing a set of notes a half-step down.

We can therefore claim that mathematics, and especially group theory, deals with symmetry in the very broad sense of "reversible (or invertible) transformations that preserve some kind of structure."

> Group theory is the study of symmetry. When an object appears symmetric, group theory can help us study it. We apply the label "symmetric" to anything that is invariant under some transformations. This can apply to geometric figures ... but also to algebraic objects like functions ....
>
> —Keith Conrad.

◇ **Exercise 3.3:** Is every group a symmetry group?

A mathematical group G is a set of objects (the group's elements) with a binary operation denoted by "+" or by "$*$" that is defined on the elements and that satisfies the following requirements.

1. Closure: for any $a, b \in G$, the sum $(a + b)$ is an element of G.
2. Associativity: any $a, b, c \in G$ we have $(a + b) + c = a + (b + c)$.
3. Identity: there exists $e \in G$ such that for all $a \in G$ $(a + e) = (e + a) = a$.
4. Inverses: for each $a \in G$, there exists a unique element $a^{-1} \in G$ such that $a + a^{-1} = a^{-1} + a = e$.
5. If the group operation is commutative, i.e., if $a + b = b + a$ for any $a, b \in G$, the group is called Abelian.

> Question: What's purple and commutes?
> Answer: An Abelian grape.

Examples of groups:

1. The set of all the integers with integer addition. The identity element is the integer 0. This is an infinite group.
2. The (finite) set of the integers 0, 1, 2,...,$m - 1$ with modulo-m addition.
3. The integers 1, 2,...,$q - 1$ for a prime q with modulo-q multiplication.
4. The set of all rotations in two dimensions under the operation: The sum of the two rotations by α and β degrees is a rotation by $\alpha + \beta$ degrees.

The set $(0, 1, 2, 3)$ with modulo-4 addition is a group denoted by $G(4)$. It obeys the addition table

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

The *order* of a group (its cardinality) is the number of elements. It is denoted by $\text{ord}(G)$. The order of $G(4)$ is 4.

The *order of a group element.* An element a of a group has an order, which is defined as the smallest positive integer n such that $a^n = e$. The order is denoted by $\text{Ord}(a) = n$ or $|a| = n$. If such an n does not exist, then a is said to have order infinity.

A simple example is the set of integers $\{0, 1, 2, 3, 4, 5\}$ with the operation of addition modulo 6. Its identity under this operation is the element 0. Choose an element $a = 5$

and examine its powers $5^1 = 5$, $5^2 = 4$, $5^3 = 3$, $5^4 = 2$, $5^5 = 1$, but $5^6 = 0$, which is why this element has order 6.

Example. The group of all positive rational numbers with arithmetic multiplication as the operation. The identity element is 1 and every element has an inverse, because there is no question of dividing by zero. It is easy to see that element 2 of the group has order infinity, because its powers under this operation are 2, 4, 8, 16, ... and none is the identity.

Example. The symmetric group S_6 of all permutations of six objects. Its elements are the permutations. Its identity is $e = \left(\begin{smallmatrix}1&2&3&4&5&6\\1&2&3&4&5&6\end{smallmatrix}\right)$, and its operation is permuting the components of an element. Consider element $a = \left(\begin{smallmatrix}1&2&3&4&5&6\\6&1&3&2&5&4\end{smallmatrix}\right)$. How many times do we have to "multiply" it by itself (i.e., to permute each of its six components in the same way) in order to bring it to the identity? We check component 1 and observe that $1 \to 6 \to 4 \to 2 \to 1$. The Mathematica statement `PermutationPower[{6, 1, 3, 2, 5, 4}, 4]` shows that this permutation is transformed to the identity after four multiplications.

Example. The dihedral group $D_{2\cdot 4}$ of all symmetries of the regular polygon with four sides (square). The elements of the group are the symmetries. The group operation $\circ$ is composition of transformations (rotations and reflections). We examine group element r_2, a 180° rotation. Obviously, two consecutive 180° rotations in the same direction bring the square to its original orientation, which is why $r_2 \circ r_2 = e$ or $\text{Ord}(r_2) = 2$.

Two interesting results can be proved from this definition of order: (1) If $a^n = e$, then the n different powers $a^0 = e$, a^1, a^2, $a^3, \ldots, a^{n-1}$ are included in the group. (2) If $\text{Ord}(a) = \infty$, then all the powers of a are different.

A *subgroup* is a subset of the elements of a group that is closed under the group's operation. A theorem by Lagrange states that if S is a subgroup of G, then $\text{ord}(S)$ divides $\text{ord}(G)$. For example, if S is the subgroup $(0, 1)$ of $G(4)$, then $\text{ord}(S) = 2$ divides $\text{ord}(G(4)) = 4$ and $G(4)$ can be partitioned into the *cosets* S and $S + 2$.

Reference [math.wrath 23] is a list of videos that illustrate many basic concepts and terms related to groups.

Note. In plain talk, we can say "if this then that," but mathematics is rigorous, which is why it is common to see the phrase "if and only if" (often shortened to "iff") in mathematical writings.

Here is a short discussion of the three main types of functions. Given two sets with elements A and B, respectively, the following terms are used:

- A general function maps each member of A to a member of B. It never maps an A to more than one B (i.e., it is not a one-to-many function), but it may map more than one A to the same B (it can be many-to-one).

- An injective function (also known as One-to-One) will not map have two or more A's to the same B. It is not many-to-one, so it is more restrictive than a general function. Also, there may be a B without a matching A.

- A surjective has at least one A that maps to every B; no B is left out.

- A bijective function (or a bijection] is one that is both injective and surjective. It provides perfect matching of the two sets (its domain and its range) in the sense that every element has a partner and no one is left out. This is referred to as a one-to-one

correspondence between the two sets. An important feature is that bijective functions have an inverse. There is always a function that maps B back to A.

Isomorphism is an important concept in group theory. Two groups are said to be isomorphic if they have the same structure. The groups may be very different, their elements, their operations, and their applications may be different, but they may still be isomorphic. It's like saying that the groups have different substances, but the same form.

Formally, two groups G_1 (with group operation $\circ$) and G_2 (with group operation $\bullet$) are isomorphic if there exists a bijection $f : G_1 \to G_2$ such that for any elements $a, b \in G_1$, $f(a \circ b) = f(a) \bullet f(b)$. In other words, if f maps a to x and b to y, then it maps the G_1 product $a \circ b$ to the G_2 product $x \bullet y$. The function f is referred to, in such a case, as an isomorphism from G_1 to G_2, which is denoted by $G_1 \cong G_2$.

Table 3.2 is an example of isomorphism. The operation of G_1 is modulo 4 addition and the $\bullet$ operation of G_2 is similar but is performed on letters. The mapping is the bijection

$$f = \begin{pmatrix} 0 & 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ e & a & b & c \end{pmatrix}.$$

$+_4$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

G_1

$\bullet$	e	a	b	c
e	e	a	b	c
a	a	b	c	e
b	b	c	e	a
c	c	e	a	b

G_2

Table 3.2: Isomorphic Groups.

Proving that two given groups are isomorphic may be challenging, but the opposite is often easy. If we can show that one of the groups has a feature that the other doesn't share, then the two are not isomorphic. Here are several examples of such features:

- One is commutative while the other one is not.
- One is cyclic and the other one is not.
- One has an element of order n (it generates a subgroup of order n) but the other one is lacking such an element.
- The two groups are finite and have different orders (different number of elements).

An important result that has to do with isomorphism is Cayley's theorem, which states that every group G is isomorphic to a subgroup of a symmetric group. An example of this kind of isomorphism is the group $Z_4 = 0, 1, 2, 3$ with addition modulo 4, listed in Table 3.2, whose elements correspond to the subgroup $\{e, (0123), (02)(13), (0321)\}$.

> There is no symmetry in nature. One eye is never exactly the same as the other.
>
> —Édouard Manet.

Groups and Symmetry

In mathematics, an important application of groups is to represent symmetry. The elements of a symmetry group are not numbers but rather symmetry operations (or transformations) on an object. The symmetry group of a geometric object is the group of all the transformations that leave the object unchanged. We say that the object is invariant under these transformations. The elements of the group are the transformations, and the group operation is composition of transformations.

A group is the perfect mathematical concept to represent symmetries, because symmetric transformations (or in short, symmetries), the transformations that preserve objects and preserve distances, must satisfy the following properties:

- A composition of symmetries is also a symmetry.
- A composition of symmetries is associative $a(bc) = (ab)c$. This is illustrated in Figure 3.3a with a pentagon. On the left side of the figure, composition P(QR) of three transformation is performed by applying R, Q, and P in this order. Transformation R is a reflection about an axis through vertex 4. Transformation Q is a reflection about a vertical axis, and transformation P is a 72° clockwise rotation. On the right side of the figure, composition (PQ)R is performed by first executing Q and P (on the top line), which produces PQ (in green), and then applying, on the bottom line, first the same PQ and then R to obtain (PQ)R. Notice that the vertex numbers are considered parts of the transformed object.
- Not perturbing an object is also a symmetry, because it preserves the object and all distances between its points. This symmetry is therefore the identity element of the symmetry group of the object.
- Any symmetry has an inverse which is also a symmetry.
- However, a symmetry group is often non-commutative, because a composition $a+b$ of symmetry transformations may be different from its reverse $b+a$. Figure 3.3b is an example. In the left part of the figure, transformations Q and P are performed in this order, to obtain the composition PQ. On the right part, the order is reversed, and it is clear that the results are different. Again, in order for the result to be obvious, the vertex numbers, which are non-symmetric, must be considered part of the object.

For simple geometric objects such as polygons and polyhedrons, it is possible to express rotations and reflections by listing the new position of each vertex of the object after the transformation. When the square of Figure 3.3c is rotated 90° clockwise, for example, vertex 1 moves to 2, 2 moves to 3, and so on, which can be expressed as the permutation

$$\begin{pmatrix} 1\,2\,3\,4 \\ 2\,3\,4\,1 \end{pmatrix}.$$

Reflections of the square about a vertical axis and about the main diagonal are expressed by

$$\begin{pmatrix} 1\,2\,3\,4 \\ 2\,1\,4\,3 \end{pmatrix}, \quad \begin{pmatrix} 1\,2\,3\,4 \\ 1\,4\,3\,2 \end{pmatrix}.$$

However, the important connection between symmetry transformations and groups becomes clear when we consider combinations of transformations.

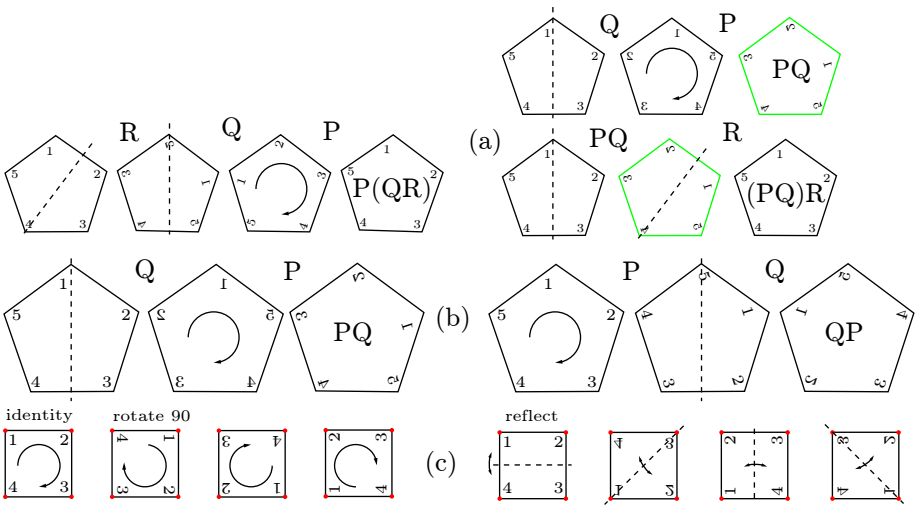


Figure 3.3: Associativity and Non-commutativity of Symmetries.

The rectangle is a simple example of this relation. It is transformed to itself after (1) a 0° rotation, (2) a reflection about a central horizontal axis, (3) a reflection about a central vertical axis, and (4) a 180° rotation. Examining any diagram of a rectangle, the following properties become clear:

1. Transformation 1 followed by any transformation t (or t followed by 1) is equivalent to just transformation t. Thus, transformation 1 is the null transformation or the identity element of the rectangle's transformation group.
2. Any of the four transformations followed by itself returns the rectangle to its original shape. Each of those transformations therefore equals its inverse.
3. Transformation 3 followed by 2 is equivalent to transformation 4.

An analysis of all the combinations of two transformations of the rectangle yields Table 3.4a. The table can be considered the definition of a symmetry group of four elements, because it specifies the group operation for the elements. A direct check verifies that element 1 (the null transformation) is the group's identity, that the operation is closed, and that it is noncommutative. This symmetry group is denoted by D_4. The D stands for *dihedral*, meaning an angle formed by two plane faces; anhedral means the opposite. In aeronautics, dihedral or anhedral is the angle—up or down, respectively—between the two wings of an aircraft.

(A dihedral group is a group whose elements correspond to a closed set of rigid motions, i.e., rotations and reflections, in the plane. The dihedral group with $2n$ elements is denoted by either D_n or D_{2n}. The group consists of n reflections, $n-1$ rotations, and the identity transformation.)

The trivial group. The identity transformation is included in any symmetry group, and is important because it is the identity element of the group. As a result, an asymmetric object, such as the letter "**J**," has a symmetry group that consists of a single element, the identity transformation e. This group is referred to as the trivial group or

the zero group. It is cyclic of order 1 and is denoted by Z_1 or C_1. Some may claim that the existence of this group implies that there is no such thing as a completely non-symmetric object, but this seems a philosophical, not a mathematical, question.

◇ **Exercise 3.4:** Explain the meaning of the phrase "Group G has no nontrivial proper subgroups."

A cyclic group, denoted by C_n, is a group generated by a single element g, the generator of the group. Every element of the group is obtained by repeatedly applying the group operation to g or to its inverse.

The complex roots of unity, discussed on Page 107, are a good example of a cyclic group. The set of complex 6th roots of unity is a cyclic group G generated by the six powers of the generator (or the primitive root)

$$z = \frac{1}{2} + \frac{\sqrt{3}}{2}i = e^{2\pi i/6}.$$

Thus $G = \{1, z, z^2, z^3, z^4, z^5\}$, where $z^6 = 1$.

Also, all cyclic groups of order 6 are isomorphic (have the same structure as) the standard cyclic group $G_6 = \{1, g, g^2, g^3, g^4, g^5\}$, where $g^6 = g^0 = 1$ and (symbol $\circ$ is the group operation) $g^i \circ g^k = g^{i+k(\mathrm{mod}6)}$.

*	1	2	3	4
1	1	2	3	4
2	2	1	4	3
3	3	4	1	2
4	4	3	2	1

(a)

*	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	1	2	3	0	6	7	5	4
2	2	3	0	1	5	4	7	6
3	3	0	1	2	7	6	4	5
4	4	7	5	6	0	2	3	1
5	5	6	4	7	2	0	1	3
6	6	4	7	5	1	3	0	2
7	7	5	6	4	3	1	2	0

(b)

*	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	1	2	3	4	0	6	7	8	9	5
2	2	3	4	0	1	7	8	9	5	6
3	3	4	0	1	2	8	9	5	6	7
4	4	0	1	2	3	9	5	6	7	8
5	5	9	8	7	6	0	4	3	2	1
6	6	5	9	8	7	1	0	4	3	2
7	7	6	5	9	8	2	1	0	4	3
8	8	7	6	5	9	3	2	1	0	4
9	9	8	7	6	5	4	3	2	1	0

(c)

Table 3.4: The D_4, D_8, and D_{10} Dihedral Symmetry Groups.

Similarly, the rhombus has limited symmetry. Its four symmetry transformations are (1) the null transformation, (2) a reflection about the line `bd`, (3) a reflection about line `ac`, and (4) a 180° rotation. An analysis of all the combinations of two of these transformations results in the same symmetry group as the rectangle. Thus, even though the rhombus and rectangle are different geometric objects and their symmetry transformations are different, we say that they have the same symmetries and we refer to them as isometric.

Intuitively, a square is more symmetric than a rectangle or a rhombus. More transformations leave it unchanged. It is easy to see that these are the four rotations by multiples of 90° and the four reflections about the vertical, horizontal, and two diagonal axes. These eight transformations can be written as the permutations

$$0=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\end{pmatrix},\quad 1=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{b}&\mathtt{c}&\mathtt{d}&\mathtt{a}\end{pmatrix},\quad 2=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{c}&\mathtt{d}&\mathtt{a}&\mathtt{b}\end{pmatrix},\quad 3=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{d}&\mathtt{a}&\mathtt{b}&\mathtt{c}\end{pmatrix},$$
$$4=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{b}&\mathtt{a}&\mathtt{d}&\mathtt{c}\end{pmatrix},\quad 5=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{d}&\mathtt{c}&\mathtt{b}&\mathtt{a}\end{pmatrix},\quad 6=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{a}&\mathtt{d}&\mathtt{c}&\mathtt{b}\end{pmatrix},\quad 7=\begin{pmatrix}\mathtt{a}&\mathtt{b}&\mathtt{c}&\mathtt{d}\\ \mathtt{c}&\mathtt{b}&\mathtt{a}&\mathtt{d}\end{pmatrix},$$

which can immediately be used to construct the symmetry dihedral group D_8 listed in Table 3.4b.

Finally, the pentagon is used to create the larger symmetry group D_{10}, because it has 10 symmetry transformations. Figure 3.3a shows that the pentagon is transformed to itself by any rotation through a multiple of 72° as well as reflections about five different axes. These ten transformations give rise to the D_{10} symmetry group of Table 3.4c.

Symmetries of the square

Figure 3.3c shows that the square has eight rigid symmetries, three rotations, four reflections, and the identity e. However, only one rotation, by 90°, and one reflection (any reflection, but we select the vertical, the one about the horizontal axis) are enough to construct the group of the square. They are denoted by r and f, respectively, and they, together with their compositions, constitute Table 3.5, the Cayley table (or multiplication table) of the square.

$*$	e	r	r^2	r^3	f	rf	r^2f	r^3f
e	e	r	r^2	r^3	f	rf	r^2f	r^3f
r	r	r^2	r^3	e	r^3f	f	rf	r^2f
r^2	r^2	r^3	e	r	r^2f	r^3f	f	rf
r^3	r^3	e	r	r^2	rf	r^2f	r^3f	f
f	f	rf	r^2f	r^3f	e	r	r^2	r^3
rf	rf	r^2f	r^3f	f	r^3	e	r	r^2
r^2f	r^2f	r^3f	f	rf	r^2	r^3	e	r
r^3f	r^3f	f	rf	r^2f	r	r^2	r^3	e

Table 3.5: The Cayley table $D_{2\cdot4}$ of the Square.

Isomorphism of groups is discussed on Page 174. The reader should notice that the two representations of the $D_{2\cdot4}$ groups listed in Table 3.4b and Table 3.5 are isomorphic.

> Every triangle is a love triangle if you love triangles.
>
> —Pythagoras.

Symmetries of the triangle

Triangles are such simple geometric figures that their symmetries can be visualized without any diagrams. Starting with the equilateral triangle. This highly-symmetric

figure transforms to itself after several rotations and reflections, transformations that are denoted here by r and f, respectively. If r denotes a 120° counterclockwise rotation, then r^2 denotes the similar 240° rotation. However, the effect of r^2 is identical to $-r$ (a 120° clockwise rotation), which is why group elements r^2 and $-r$ are inverses of each other. Similarly, transformation r^3, with is a full 360° rotation equals the group identity, denoted by e. The rotation part of the symmetry group of this triangle can be summarized by either the short list

$$e, r, r^2 = -r, r^3 = e,$$

or by writing the group elements in terms of the vertex numbers, using cycle notation

$$e = \begin{pmatrix} 1,2,3 \\ 1,2,3 \end{pmatrix} = (1)(2)(3),\ r = \begin{pmatrix} 1,2,3 \\ 3,1,2 \end{pmatrix} = (1,3,2),\ r^2 = \begin{pmatrix} 1,2,3 \\ 2,3,1 \end{pmatrix} = (1,2,3).$$

A symmetric reflection f_i of the equilateral triangle is a flip about the axis that passes through vertex i. There are three such reflections and it is easy to visualize them and convince ourselves that $f_i^2 = e$ and $f_i = -f_i$ is its own inverse. Thus, reflections add more cycles to the symmetry group of the equilateral triangle (the dihedral $D_{2\cdot 3}$ group), which is summarized in Table 3.6.

$$f = \begin{pmatrix} 1,2,3 \\ 1,3,2 \end{pmatrix} = (1)(2,3),\ rf = \begin{pmatrix} 1,2,3 \\ 2,1,3 \end{pmatrix} = (1,2)(3),\ r^2 f = \begin{pmatrix} 1,2,3 \\ 3,2,1 \end{pmatrix} = (1,3)(2).$$

	e	r	r^2	f_1	f_2	f_3
e	e	r	r^2	f_1	f_2	f_3
r	r	r^2	e	f_3	f_1	f_2
r^2	r^2	e	r	f_2	f_3	f_1
f_1	f_1	f_2	f_3	e	r	r^2
f_2	f_2	f_3	f_1	r^2	e	r
f_3	f_3	f_1	f_2	r	r^2	e

Table 3.6: Symmetry Group of the Equilateral Triangle.

The isosceles triangle features lower symmetry. It only has one reflection in addition to the identity, so its symmetry group consists of only e and f. A scalene triangle has no rotation or reflection symmetries.

Symmetric groups and cycle notation

Permutation groups are a special case that shows up in connection with symmetry. The elements of a permutation group are permutations of a given set M of symbols, and the group operation of the permutation group is composition of permutations (which are thought of as bijective functions from the set M to itself). The connection between permutation groups and symmetry is the group action, the way in which the elements of a permutation group permute the symbols of set M. This is easily grasped with examples.

The group of all permutations of a set M is the symmetric group of M, often written as Sym(M). A permutation group is therefore a subgroup of some symmetric group. If M is the set $\{1, 2, \ldots, n\}$, then Sym(M) is normally denoted by S_n, and may be called the symmetric group on n symbols (not to be confused with symmetry groups). The well-known Cayley's theorem states that every group is isomorphic to some permutation group.

Some permutation groups are also symmetric groups, which is not the same as symmetry groups. Symmetric groups are denoted by S_n where n is the number of symbols in the permutation. The elements of the group itself are the $n!$ permutations, not the symbols that are permuted. Thus, for example, symmetric group S_3 is the set of six permutations of three basic elements. It can be written in cycle notation (see below) as

$$S_3 = \left\{ \begin{bmatrix} 1,2,3 \\ 1,2,3 \end{bmatrix}, \begin{bmatrix} 1,2,3 \\ 1,3,2 \end{bmatrix}, \begin{bmatrix} 1,2,3 \\ 3,2,1 \end{bmatrix}, \begin{bmatrix} 1,2,3 \\ 2,1,3 \end{bmatrix}, \begin{bmatrix} 2,3,1 \\ 1,2,3 \end{bmatrix}, \begin{bmatrix} 1,2,3 \\ 3,1,2 \end{bmatrix}, \right\} \qquad (3.1).$$

Notice that the order `S_3` of this group is the same as that of D_3 (the dihedral group of order 3 or the group of the symmetries of an equilateral triangle), but this is not always the case. The order of S_4 is $4! = 24$, which is different from $|D_4| = 8$. Every element of D_4 is a permutation, and is also an element of S_4,

The notation used in Equation (3.1) is cumbersome, which is why the cycle notation of permutations is sued in practice. Using this notation, group S_3 is written

$$S_3 = \{(1,2,3),\ (1)(2,3),\ (1,3)(2),\ (1,2)(3),\ (2,1,3),\ (1,3,2)\},$$

where the cycle $(1)(2,3)$ means $1 \to 1,\ 2 \to 3,\ 3 \to 2$. Often, in cycles such as $(1,2)(3)$, and $(1)(2,3)$, an author may omit the short cycle with a single element and write only $(1,2)$ or $(2,3)$.

When composing symmetry transformations, we end up with a cycle for each transformation and then combine them in a simple way as follows. Given the transformations Q, with cycle $(1,3,5,2)$, and P, with cycle $(2,4,3)$, if P is applied first, then the composition is written QP and it corresponds to the combined cycle $(1,3,5,2)(2,4,3)$. Going from right to left, we notice that 1 is first permuted to 1 and then to 3, so in the combined cycle 1 goes to 3. Similarly, 2 first goes to 4, and later, 4 goes to itself, so when these are combined, 2 goes to 4. Similarly $3 \to 2 \to 1$ and $1 \to 1 \to 3$. The combined cycle is now $(1,3)(2,4,5)$.

◇ **Exercise 3.5:** Try your hand with the two cycles left $(2,9,7,4)$ and right $(1,6,4)$.

◇ **Exercise 3.6:** Write the cycle notation of dihedral group $D_8 = D_{2\cdot 4}$.

Symmetry in three dimensions can be richer and much more complicated than its two-dimensional equivalent. It is also much harder to analyze and visualize. A cube is a good example. This simple geometric object can be rotated and reflected in many ways and it features high symmetry. We concentrate on dice, which are cubes with their faces labeled and therefore distinct. The figure shows three views of one of the



eight corners of a die, the corner adjacent to the faces labeled 1,
2, and 3. Each view can be obtained from any of the other ones by a rotation, for example, a rotation that places one of the faces on top. This is why these three rotations are part of the symmetry group of the die.

Having eight corners, the group therefore features 24 rotation symmetries. However, the particular labeling of the die gives rise to another group of 24 rotation symmetries. A careful examination of the dice in the figure will convince the reader that the faces seen at the one/two/three corner are arranged counter-clockwise. In principle, it is possible to have dice where the same faces are arranged clockwise, which results in 24 additional rotation symmetries of the die, for a total of 48.

(The opposite faces of a die are traditionally arranged to always sum to seven. This gives two possible mirror image arrangements in which the numbers 1, 2, and 3 may be arranged in a clockwise or counterclockwise order about a corner. Figure 3.7 illustrates this chirality of dice. The cube faces may be placed counterclockwise (left) or clockwise (right).)

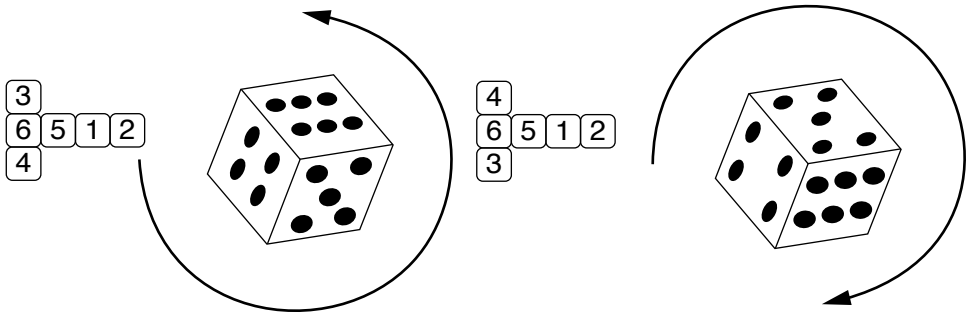


Figure 3.7: The Chirality of Dice.

◇ **Exercise 3.7:** The six faces of a standard die are numbered 1 through 6. The arithmetic mean of these numbers is $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$), which is what we expect when a single die is thrown many times. Throwing two dice results in 2 through 12, with a mean of 7. Design a numbering system for a six-sided die that produces the same means.

It is clear that symmetries in plane (two-dimensional) figures can be quite complicated and may be difficult to visualize. Symmetries in three-dimensional geometric figures can be much more complex and harder to see. An important difference to keep in mind is the behavior of reflections. A two-dimensional figure has infinite thickness and therefore no front and back. When reflecting such a figure, we see the same side which is both the front and the back. This is why certain two-dimensional reflections are symmetry operations. A three-dimensional figure does have a front and a back, so when it is reflected (flipped) about any axis, we see its back. If its back and front are different (have different colors or patterns) the reflection is not a symmetry transformation. See also Exercise 3.8.

◇ **Exercise 3.8:** Given the red, two-dimensional, two-part asymmetric figure of Figure Ans.9, show how to transform it to itself, ending up in the same location and using just rotations and reflections, no translation.

Symmetries of the cube

A cube is a polyhedron, so it consists of surfaces, which are bounded by edges, which in turn are delimited by corner points (vertices). Thus, any symmetry of a cube is a special permutation of its eight corners, a permutation that transforms edges to edges. The pair of corners of an edge are moved by a symmetry transformation to become the pair of corners of some edge. A cube has 48 symmetries of which 24 are rotations and can be performed physically on a real cube. The other 24 are reflections or rotary reflections, and can be visualized with the help of diagrams or videos.

The 24 rotational symmetries of the cube can be counted and understood in three steps as follows:

1. Rotation about the faces. Imagine an axis through the center of a face of the cube and its opposing face. Turning the cube about this axis by 90°, 180°, and 270°, are symmetry transformations. The six faces of the cube make three pairs of opposing faces, which produces $3 \times 3 = 9$ symmetry rotations. This is particularly easy to see if the faces are numbered or painted different colors.

2. Rotations about the edges. A cube has 12 edges which can be organized in six pairs of opposing edges. Imagine an axis from the center of an edge, through the center of the cube, to the center of the opposing edge. A 180° rotation abut this axis produces a symmetric transformation. These make six more symmetry rotations of the cube.

3. Finally, there are rotations about the eight corner points. These can also be organized in four sets of two opposing points each, and rotations by 120° in one direction and 120° in the opposite direction—about the axis going from a point, through the cube center, to the opposite point—produce two more symmetry transformation for each pair of points, for a total of $2 \times 4 = 8$ rotations.

This brings the total of symmetric cube rotations to $9 + 6 + 8 = 23$, which together with the ubiquitous identity transformation, brings the total to 24.

We turn to the reflections. This kind of symmetry transformation is performed about a plane and cannot be done on a real, solid cube. Figure 3.8 shows three planes centered in the cube, each between two pairs of faces. These give rise to three reflections. The figure also shows six more planes, each located between two opposing edges and each allows for a reflection. The total number of reflections is therefore nine.

Is that all? No, because the 24 rotations form a subgroup of symmetries, and the nine reflections indicate that the order of the full group must be greater than 24. Lagrange's theorem implies that the total number of symmetries must be a multiple of 24. We know that it is strictly larger than 24, and we later show that it cannot be higher than 48, so there must be 15 more symmetry transformations of the cube. Those transformations are referred to as rotary (or turn) reflections. They either start with a reflection which is followed by a rotation, or start with a rotation and follows it by a reflection. The result of a rotary reflection cannot be achieved by a rotation alone. The 15 rotary reflections of the cube fall into three categories as follows:

1. Choose any pair of opposite faces, reflect the cube about this pair, and then rotate it about the axis connecting the two faces, by 90° in one direction and 90° in

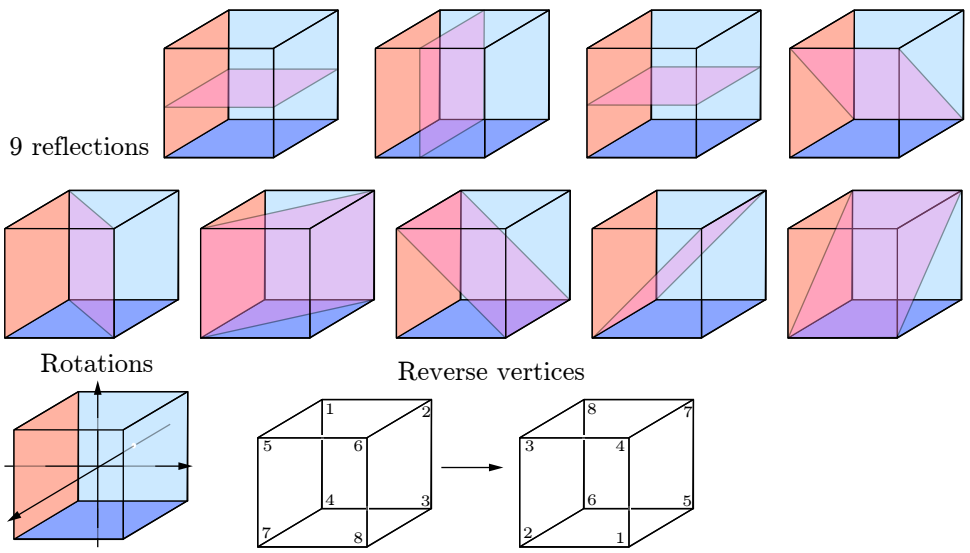


Figure 3.8: Symmetry Operations in a Cube.

the opposite direction. The results of these transformations cannot be obtained by a rotation alone. There are three pairs of opposite faces, which contributes six symmetries to the symmetry group of the cube.

2. Select a pair of opposite vertices, reflect the cube about the axis that connects them (a diagonal that passes through the center of the cube), and rotate the result about that axis 60° in one direction and 60° in the opposite direction. The four diagonals contribute eight additional rotary reflection symmetries to the group.

3. Finally, reflect every vertex about the center of the cube. No need for an extra rotation. Shown in Figure 3.8, this includes the last symmetry transformation in our symmetry group of the cube.

The total of the rotary reflections is therefore $6 + 8 + 1 = 15$, which completes the order of the symmetry group to $1 + 23 + 9 + 15 = 48$.

Something to think about. Given a solid cube, any transformation on it, symmetric or not, moves many points and may leave some points unchanged. A rotation leaves unchanged all the points on its axis. A reflection leaves unchanged all the points on the reflection plane. A rotoreflection leaves only the center point unchanged.

Can there be more than 48 symmetric cube transformations? A simple answer (although not a formal proof) is to look at a corner of a cube. You see three edges emanating from it, leading to three other corners. These three corners can be arranged in six ways, since each arrangement is a permutation of the corners. There are eight corners with six permutations each, so the maximum number of transformations is $6 \times 8 = 48$.

◇ **Exercise 3.9:** What is the multiplication table of the symmetry group of the cube

Symmetries of the tetrahedron

In everyday talk, a tetrahedron is a triangular-based pyramid. It is therefore surprising to learn the this simple, familiar object (the simplest Platonic solid) features rich symmetry, some of which is difficult to visualize. Even a cursory look at a real pyramid suggests that it has rotational symmetry, and it doesn't take much imagination to guess some of its reflection symmetries, but the full group of tetrahedron symmetries takes time to construct. While we are at it, we should mention that the tetrahedron does not have a center, which is why it does not have inverse symmetry (where every vertex is swapped with its opposite).

We start with the rotations. Figure 3.9 depicts the 12 symmetry rotations, and it is clear that most are about a vertex. Each row of the figure shows two 120° rotations as seen from a vertex. The left column includes the clockwise rotations and the right column has the counterclockwise rotations. They total eight rotations. Three more 180° rotations are shown in the center column. They are done about an axis that passes from the midpoint of an edge to the midpoint of the opposite edge (Figure 3.10a). The six edges of the tetrahedron provide for three pairs of opposing edges. Together with the identity transformation, the group of rotations is of order 12. The permutations of vertex numbers are listed in cycle notation.

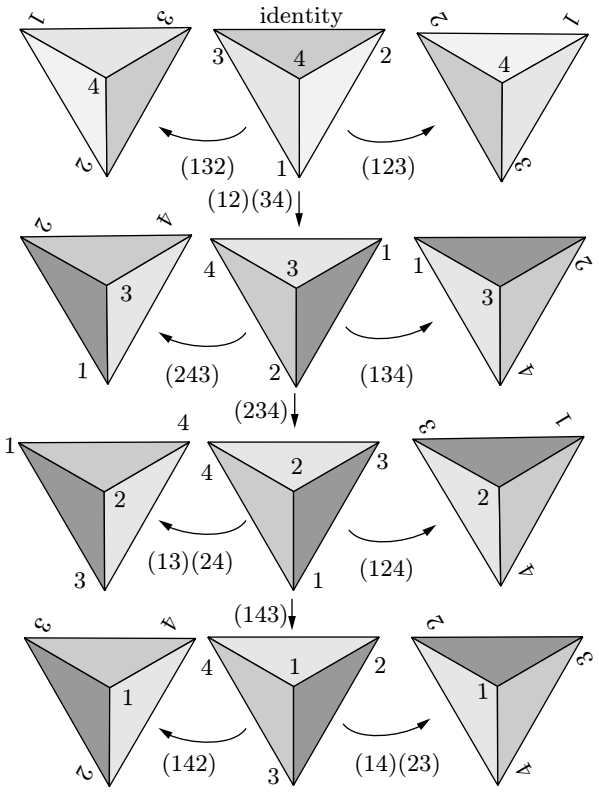


Figure 3.9: 12 Symmetry Rotations of the Tetrahedron.

Reflections are more difficult to visualize. Figure 3.10b Shows one of six reflection symmetries of the tetrahedron. Each reflection is about a plane that goes from an edge to the midpoint of the opposite edge. There are also six rotoreflections by 60° (part (c) of the figure) for a total of 24 symmetries.

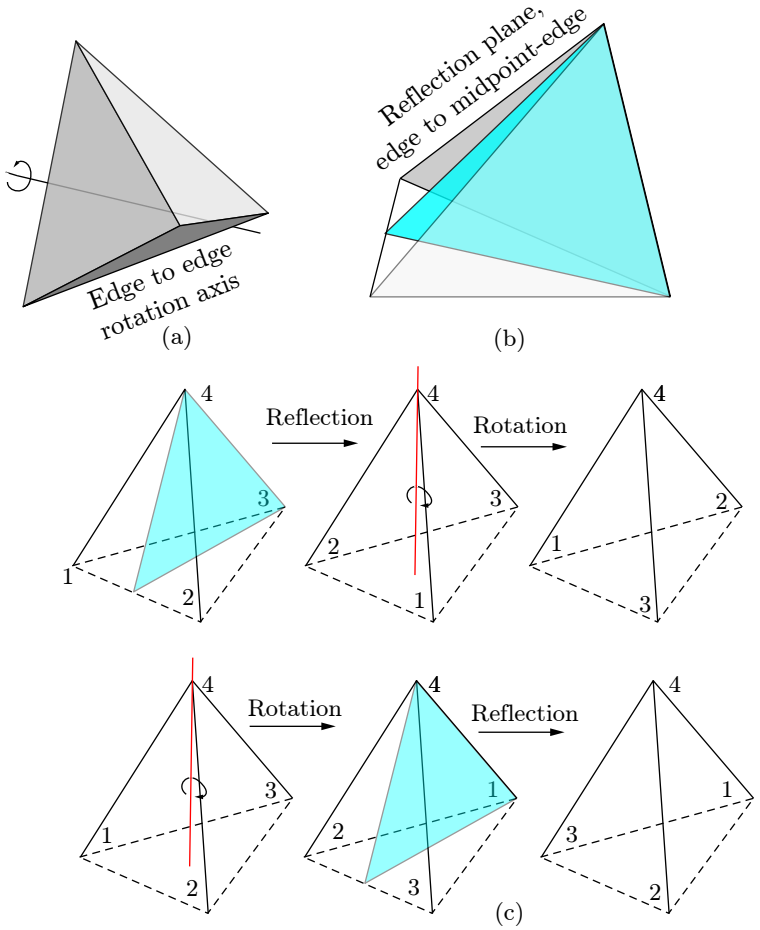


Figure 3.10: Reflections and Rotoreflections of the Tetrahedron.

A careful analysis of the permutations of the vertices of the tetrahedron shows that there is exactly one symmetry transformation for each permutation. The set of 24 tetrahedron symmetries is therefore isomorphic (has the same structure as) group S_4, the symmetric group of permutations of four objects. (The word "symmetric" here means the full set of $4! = 24$ permutations.)

Table 3.11 lists all 24 symmetries in cycle notation, where R_1 means a rotation about vertex 1, and F means reflection. A curious fact is that reflections (2,1,3,4),

(1,3,2,4), and (1,2,4,3) are a basis of this group of symmetries, because each of the 24 symmetries can be expressed a product of two of these three. Here are some examples.

$$(2,1,3,4)\circ(1,3,2,4)=(2,3,1,4),\quad (2,3,1,4)\circ(2,1,3,4)=(3,2,1,4),$$
$$(2,3,1,4)\circ(2,3,1,4)=(3,1,2,4),\quad (2,1,3,4)\circ(1,2,4,3)=(2,1,4,3)$$
$$(1,3,2,4)\circ(1,2,4,3)=(1,3,4,2).$$

$(1,2,3,4)$	E	$(1,3,4,2)$	R_1	$(1,4,2,3)$	R_1
$(4,2,1,3)$	R_2	$(3,2,4,1)$	R_2	$(2,4,3,1)$	R_3
$(4,1,3,2)$	R_4	$(3,1,2,4)$	R_4	$(2,3,1,4)$	F
$(2,1,4,3)$	F	$(3,4,1,2)$	F	$(4,3,2,1)$	F
$(1,2,4,3)$	F	$(1,3,2,4)$	F	$(2,1,3,4)$	F
$(1,4,3,2)$	F	$(4,1,2,3)$	F	$(2,4,1,3)$	F
$(3,1,4,2)$	F	$(3,2,1,4)$	F	$(2,3,4,1)$	F
$(4,3,1,2)$	F	$(4,2,3,1)$	F	$(3,4,2,1)$	F

Table 3.11: The 24 Symmetries of the Tetrahedron.

◇ **Exercise 3.10:** Why does a triangular-based pyramid have more symmetries than a square-based pyramid?

Isometry, also called rigid motion, is a transformation that preserves distances. It can be shown that there are only four plane isometries: translation, reflection, rotation, and glide reflection. Together, the four are known as the basic rigid motions of the plane.

Glide reflection. A glide reflection (or a miracle, a term coined by John Conway), is a composition of a reflection and a translation along the line of the reflection. This symmetry is commutative; a reflection followed by a translation is the same as a translation followed by a reflection. Figure 3.12 is an example. Notice that a glide reflection can be seen as an improper rotation with an infinite radius (a circle with infinite radius is a straight line)

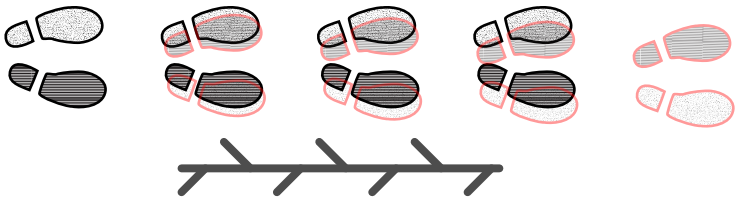


Figure 3.12: Glide Reflection Symmetry.

In addition to the isometries, the following symmetric transformations exist.

Gyration. Consider an object that has both rotation and reflection symmetries. Normally, any point that is the center of a symmetric rotation is also located on some

line of reflection. However, the Roman letters N S Z have a 180° rotation symmetry about their center but no reflection symmetry. This is an example of gyration symmetry, a term coined by John Conway.

Figure 3.13 illustrates a more complex gyration. The red point is the center of a 120° symmetric rotation, but is not on any line of reflection. The nearest reflection line (in green) is shown, together with another center point of rotation. The red point is the center of a gyration symmetry.

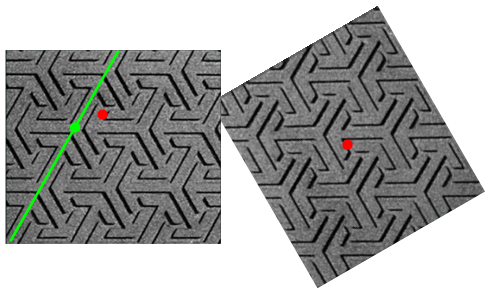


Figure 3.13: Gyration Symmetry.

◇ **Exercise 3.11:** A summary. Identify the symmetries of an infinite wall of bricks.

Mathematicians like to extend and generalize new concepts and term, so they have also assigned an order to a gyration point. The order of such a point is n if the smallest angle of any rotation about it is $2\pi/n$.

Wikipedia offers the following formal definition of gyration. In geometry, a gyration is a rotation in a discrete subgroup of symmetries of the Euclidean plane such that the subgroup does not also contain a reflection symmetry whose axis passes through the center of rotational symmetry.

Some people may prefer the following, more visual, definition: A gyration point is a rotation point that does not lie on a mirror.

Improper rotation, also called rotoreflection or rotary reflection, is an isometry in Euclidean space that is a combination of a rotation about an axis and a reflection in a plane perpendicular to that axis.

> [John] Conway's office was a shrine to symmetry.
> —Marcus du Sautoy, *Symmetry, Journey into the Patterns of Nature.*

Kaleidoscopic Points

A kaleidoscopic point is the center point of reflections and often also rotations in an image. There may be several kaleidoscopic points in an image. This insert talks about kaleidoscopes and how they generate their images.

A kaleidoscope, a simple optical instrument with mirrors, is a familiar object, normally used as a toy. It uses two or three mirrors to create complex symmetrical patterns out of a simple input pattern which is reflected multiple times. The input is colored

pieces of glass, ribbon, confetti, glitter, feathers, flower, beads, or disks, which are contained inside the instrument and change positions as it is rotated. In a teleidoscope, a near cousin of the kaleidoscope, the pattern is created outside the instrument.

The kaleidoscope was invented by David Brewster in 1817 and was named after the Greek words καλός (*kalos*, "beauty"), εἶδος (*eidos*, "that which is seen"), and σκοπέω (*skopeo*, "to look to").

> Every little development that arises is like the shake you give to a kaleidoscope—the thing changes entirely in aspect.
>
> —Agatha Christie, *The Murder of Roger Ackroyd*, 1926.

Most kaleidoscopes have three mirrors, tilted to one another at 60°, but several variations are possible, some even proposed by Brewster himself. They are listed here and are recommended to anyone who would like to experiment with a home-made kaleidoscope:

- Two mirrors, tilted to each other at 60°, with a black rectangle instead of a third mirror.

The following can be referred to as Polycentral Kaleidoscopes.

- Three mirrors at 90° and a black fourth side. This creates an infinite pattern of equilateral triangles.
- Three mirrors at 90°, 45°, and 45°. This creates an asymmetric pattern.
- Three mirrors at 90°, 60°, and 30°. The pattern viewed consists of 31 reflected images of the aperture of the kaleidoscope.
- Four mirrors. This generates a pattern with symmetry based on squares and rectangles.
- A regular dodecahedron. This is a Platonic solid, a polyhedron, with 12 regular pentagons as faces. The mirrors must be trimmed at each corner, and are glued or taped together. The trimming creates small triangular holes around the entire kaleidoscope, through which the user can peep and also shine light. The patterns viewed by the user are created by the light coming through the holes and are three-dimensional and very complex.
- In a similar way, three-dimensional kaleidoscopes can be made from the other Platonic solids, but I haven't discovered any.

Figure 3.14 illustrates typical examples of kaleidoscopic images with their reflection axes. They also have rotation symmetries which are easy to located.

A home-made kaleidoscope can be based on real mirrors, cut to size by a professional, or on sheets of aluminum foil, parts of old CDs or DVDs, or even silver-coated wrapping paper.

Because of the use of mirrors, kaleidoscopic patterns are rich in reflections, which is why patterns whose symmetries are dominated by reflections are called kaleidoscopic, and why points of intersection of reflections lines are referred to as kaleidoscopic points.

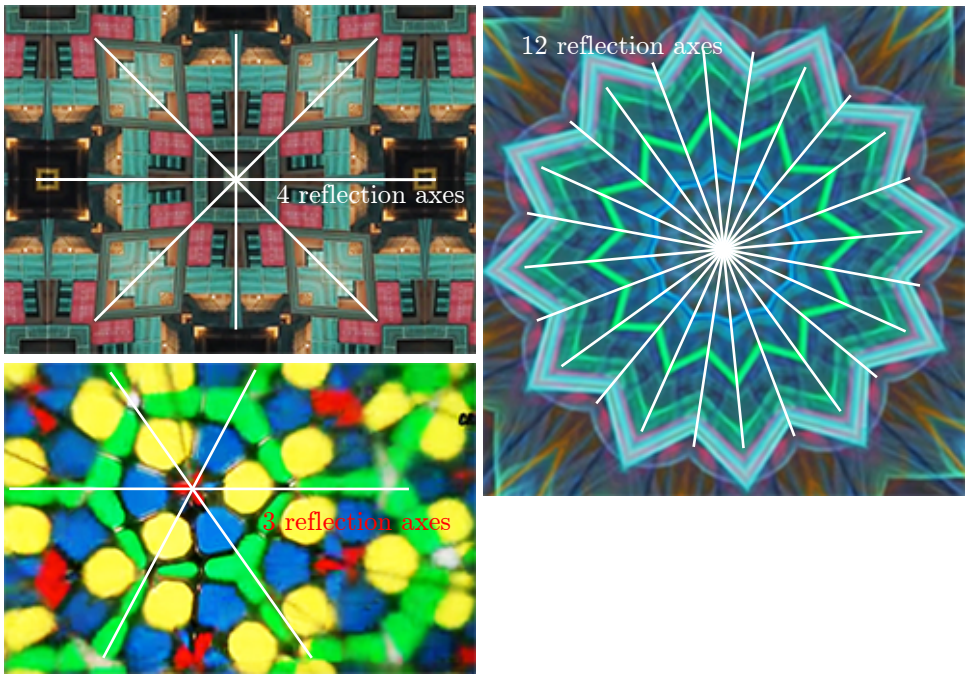


Figure 3.14: Kaleidoscopic Images and Reflection Axes.

3.3 Orbifold Notation

Once we grasp the concept of a mathematical group, we naturally want to "visualize" the group, i.e., to see a diagram, a table, or an expression that summarizes or enumerates the essence of the group or that lists all its elements (if it is not too big). Several such methods have been proposed over the years and are in use (see Appendix A of reference [Conway 08]). Important notations are the Coxeter, Schoenflies, Hermann-Mauguin, and IUC (International Union of Crystallography). This section describes the basics of the orbifold notation, often referred to as the orbifold signature. This notation consists of a compact string of digits and a few symbols that summarize the symmetry features of an object. General references for this section are [Conway 08], [du Sautoy 09], and [Conway et al. 91].

The orbifold notation was developed by William Thurston and promoted by John Conway. It is used for representing types of symmetry groups of two- and three-dimensional objects. The term "orbifold" is an abbreviation of "orbit-manifold," It has to do with folding symmetric patterns, and can best be visualized by watching the videos in references such as [Gould 20] and [GregSmith 15].

> This terminology should not be blamed on me. It was obtained by a democratic process in my course of 1976–77. An orbifold is something with many folds; unfortunately, the word "manifold" already has a different definition. I tried "foldamani," which was quickly displaced by the suggestion of "manifolded." After two months of patiently saying "no, not a manifold, a manifoldead," we held a vote, and "orbifold" won.
>
> —Thurston explaining the origin of "orbifold." Section 13.2 of [Thurston 80].

The principle of any notation of symmetry groups is to define various types of symmetries (rotation, reflection, and so on), to assign symbols to those types, and to end up with a compact "signature" for any pattern with symmetries, such that patterns with the same signatures would have the same symmetries even though they may look very different when viewed. Conversely, patterns that look similar may have different symmetries and different signatures.

In order for such a notation to be perfect, it must restrict itself to a class of patterns, such as wallpaper patterns, frieze patterns, or three-dimensional patterns, and has to identify all the symmetry types of the patterns in the class. As an example, if we ignore rotation and translation symmetries and identify only reflection (mirror) symmetries, we will have signatures of two types, type X may be patterns with mirror symmetries and type Y may be patterns without mirror symmetries; an imperfect classification that ignores important types of symmetry.

Those parts of the orbifold notation that are described here are limited to wallpaper patterns, but the orbifold notation includes symbols for other classes of patterns, especially frieze- and three-dimensional patterns. A wallpaper pattern is an infinite, two-dimensional abstract mathematical object that covers the plane by repeating a motif indefinitely, while keeping certain isometries unchanged. A frieze pattern is similar, but repeats in one direction only. Frieze patterns are common in architecture and decorative art. They have been classified into seven types according to their symmetries, but will not be discussed here.

The orbifold notation can be used for symmetry groups of patterns that include one or more of the following symmetries:

- Reflection through a line (in 2D) or a plane (in 3D). The order of a reflection is the number of different reflection lines or planes.
- Translation by a vector.
- Gyration of finite order about a point. The order n of a gyration of θ radians is defined as $n = 2\pi/\theta$. Gyration symmetries (Page 187) are denoted by blue digits in the examples that follow.
- Infinite rotation around a line in 3-space.
- Glide-reflection, i.e. reflection followed by a translation.

An orbifold signature may include some or all of the following symbols:

- An asterisk "*" indicates an object with reflection symmetries. The selection axes are denoted by red digits as illustrated by the examples below.
- A red point (•) indicates that all the symmetries of a pattern fix a single point.

- The × symbol, which is called a miracle (short for "mirrorless crossing"). It indicates a part or parts of the pattern that feature a glide reflection symmetry that doesn't cross any reflection axes.

- Many wallpaper patterns feature translation in addition to other types of symmetries. If translation is the only kind of symmetry in a pattern, its orbifold signature is 0, a symbol derived from the term "wonderful wandering," or "wonder-ring."

We are now ready for the details of the orbifold notation, presenting it with examples. First, reflections, which are referred to as kaleidoscopes. The previous discussion of kaleidoscopic points explains the connection between reflections (mirror symmetry) and kaleidoscopic symmetry, and shows why the latter term is more general. Figure 3.15a shows a heart, a simple plane object with a single reflection line (in red). Its orbifold signature is a single star, denoted by ∗ or ∗1.

(In general, orbifold signatures can be written in black, but it is convenient to specify kaleidoscopes and miracles in red and gyrations and wonders in blue. Blue is used for symmetries that preserve the true orientation of a fundamental region of the pattern being examined. Red is used for kaleidoscopes and miracles, since these reflect.)

The star denotes kaleidoscopic symmetry, and in our example of Figure 3.15a, where there are no other symmetries, the ∗ (or the equivalent ∗1) is the entire signature.

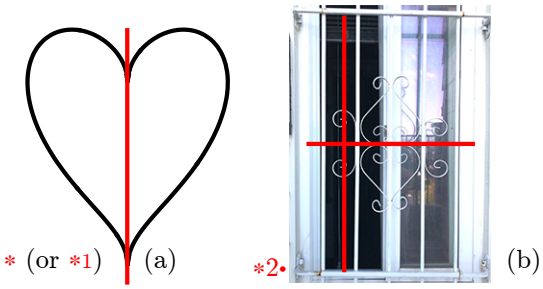


Figure 3.15: Orbifold signatures ∗ (or ∗1) and ∗2•.

Part (b) of Figure 3.15 has signature ∗2•, pronounced "star two point symmetry." The star indicates kaleidoscopic symmetry and the 2 is the period of this symmetry, indicating two reflection axes (mirror lines). The point (•) indicates that all the reflection axes pass through a point (otherwise, they must be parallel). In general, a point at the end of a signature indicates that all the symmetries, not just reflections, fix a point in the pattern/object.

◇ **Exercise 3.12:** What letter-string of the Roman alphabet has a ∗2• signature?

Images with more reflection axes can similarly have signatures ∗3•, ∗4•, and so on, as shown in Figure 3.16. It is obvious that these objects also have rotation symmetries, but these are not included in the signature, because every intersection of reflection lines determines a rotation point. This little-known fact is discussed on Page 218 and is illustrated by Figure 3.46. Any rotation points located on reflection lines are not considered independent symmetry features and are not included in orbifold signatures.

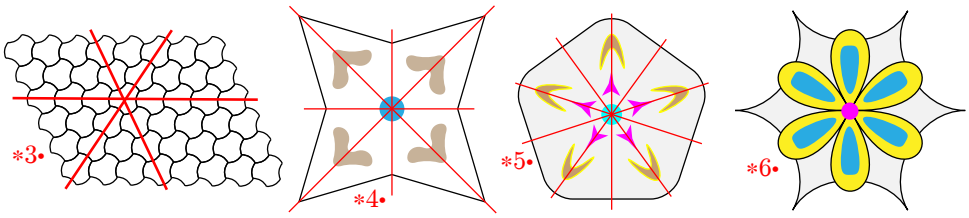
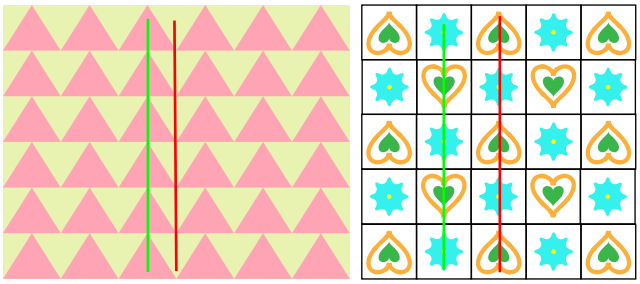


Figure 3.16: Several Different Reflection Axes.

For rotation symmetries, the signatures include only gyration points, which are rotation points not located on any reflection lines. Figure 3.13 is an example of such a point.

A similar, but slightly different, symmetry is illustrated in Figure 3.17. The patterns have reflection symmetries about two lines/axes (green and red) that are parallel, but also different (placed at different points in the symmetry), which is why the signatures of these patterns are ** (star star and no point). Also, since * is the same as *1, signature ** is the same as *1*1. It is understood that our symmetric patterns are infinite and feature infinite reflection axes, but the green and red axes are positioned at different places in the symmetry, which is why they are considered different. (Rule: if two features are located at points that are symmetrically identical, only one is counted.)



Signature **

Figure 3.17: Patterns with Signatures **.

Figure 3.18 is an example of slightly more complex symmetry. There are two sets of two reflection axes each, and the axes of each set intersect at a point. Each of those points is also the center of rotation, but these rotations are not included in the signature as explained earlier. On the other hand, there is a gyration point (of order 2), marked in blue, which should be included. All this implies the signature 2*22. Also notice that we can choose either the two intersection points marked 1 or the two marked 2. The two sets of reflection axes may be the red-green and the blue-cyan, but they may also be the red-blue and green-cyan.

A more complex kaleidoscopic symmetry is illustrated in Figure 3.19. At first, the figure seems simple, consisting of hexagons only, but careful examination reveals three centers of reflection, with orders 6, 3, and 2, respectively. The signature is therefore *632,

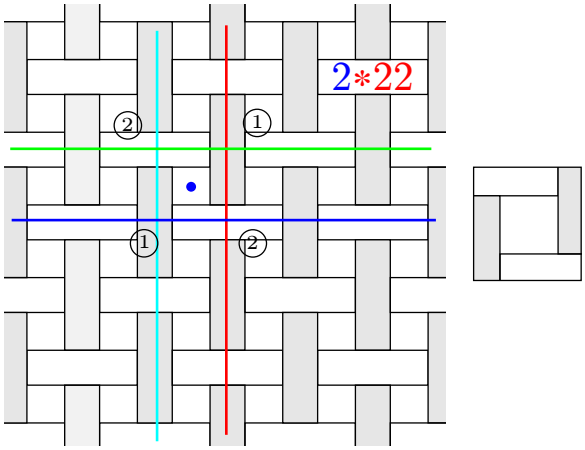


Figure 3.18: A Pattern with Signature 2∗22.

which can also be written as cyclic permutations of 632 and of its reverse 236. Having just hexagons, it is no wonder that this pattern also exhibits rich rotational symmetry, but its rotation centers coincide with the kaleidoscopic points, and are therefore not gyrations.

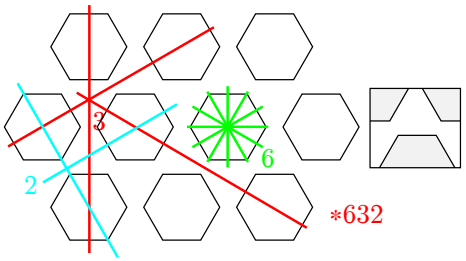


Figure 3.19: A Pattern with Signature ∗632.

◇ **Exercise 3.13:** Find patterns with signatures ∗442 and ∗333. Discuss any gyration and rotation symmetries that they may have.

Fundamental regions. In a pattern, find a region surrounded by mirror lines that doesn't itself contain any mirror lines. Such a fundamental region acts like a stamp whose repeated applications generate the entire pattern. The pattern can also be generated by placing mirrors at all the mirror lines of the fundamental region, much as is demonstrated in reference [Gould 20] or in a kaleidoscope, which is why the term "kaleidoscopic" is more general than "reflection." However, physical kaleidoscopes are limited to rectangles, equilateral triangles, and 90°, 60°, 30°, and 90°, 45°, 45° triangles, while our fundamental regions may have any polygonal shapes. Several of the figures in this section also include a fundamental region.

Gyration symmetries are indicated in the orbifold notation with a (blue) number to the left of the star (if any), indicating the order ($2\pi/\theta$) of the rotation. The star is included only if there are also reflection symmetries. The objects of Figure 3.20 do not feature any reflection symmetries, so their signatures don't have a star. They are simply 32∗22, 4, and 5. In contrast, the patterns of Figure 3.16 have rotation, but not gyration, symmetries in addition to the reflections, which is why their signatures do not specify any rotational symmetries, and are all red.



Figure 3.20: Patterns with Gyration Symmetries.

More gyration symmetry is illustrated in Figure 3.21. In the pattern on the left, the centers of rotation, all at 180° (order 2) are at the middles of the large, middle, and small diamond shapes. There being no reflection symmetries, the signature is simply 2222. In the pattern on the right (compare with Figure 3.18), there is one gyration point, marked in blue, and two sets of reflection axes, for a signature 2∗22.

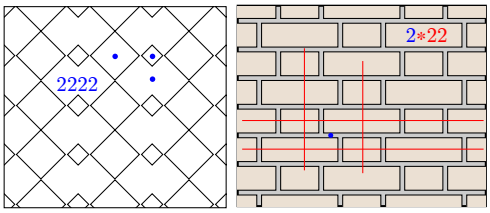


Figure 3.21: Patterns with Signatures 2222 and 2∗22.

◇ **Exercise 3.14:** Try to guess a pattern with a 3∗3 signature.

The next important feature of symmetrical patterns is dubbed a "miracle," a term which stands for "mirrorless crossing." In the pattern, a miracle is denoted by two red dots and a dotted curve connecting them, while in the signature, a miracle is denoted by an ×. A miracle occurs when a fundamental region of the pattern "covers" other regions of the pattern perfectly by shifting and reflecting it, but without crossing reflection

axes. In other words, it is a glide reflection that doesn't cross any reflection axes. A miracle can therefore be viewed as a symmetry feature that may convert image features (motifs) between left-handedness and right-handedness (something that can be done by a reflection but not by a rotation) and can do so locally, i.e., without the "help" of the pattern's "main" reflection axes.

Figure 3.22 shows examples of $\times$ and $\times\times$ (miracle and double miracle) signatures. Notice that the latter pattern has no reflection symmetries at all, which is why it is possible to find regions in this pattern that do not cross any reflection axes.

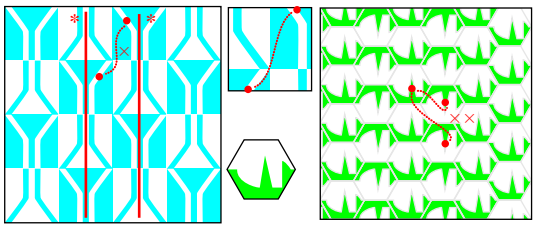


Figure 3.22: Patterns with Signatures $*\times$ and $\times\times$.

◇ **Exercise 3.15:** Find a pattern with a double miracle $\times\times$ signature.

Translations (∘ notation). In any symmetry transformation that involves rotation and/or reflection, there must be in the (infinite) pattern at least one point that does not change location during the transformation. Such a point can be considered the center of mass of the pattern. Translation is different from rotation and reflection in this sense, being a transformation where all the points of a pattern move in unison. The entire pattern moves rigidly in a certain direction, without rotations or reflections. A test for the presence of translation symmetry in a pattern is to identify the fundamental region of the pattern and check if it can be repeated to generate the entire image without having to reflect or rotate copies of the region. A miracle is similar, except that the fundamental region must be rotated or reflected to fit parts of the pattern.

The patterns shown here so far are commonly referred to as wallpaper patterns; they are not the only possible patterns and they are distinguished from other patterns by the following features:

- Wallpaper patterns have translation symmetry. Such a pattern can be shifted in at least two directions. In any direction, there is always a smallest unit of translation.

- Wallpaper patterns may also feature reflection, rotation, and glide reflection symmetries.

Given patterns such as those in Figure 3.23, with translations, but with no rotations, reflections, or glide reflections, their signatures are a single ∘, a symbol that comes from the term "wonderful wandering," or "wonder-ring."

Summary

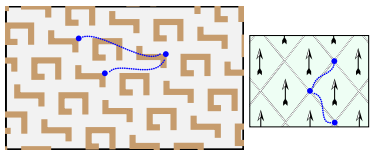


Figure 3.23: Patterns with Signature 0.

Given a wallpaper pattern (i.e., repeating and two-dimensional), the following steps can be used as guidelines for determining its orbifold signature:

- Locate any mirror (reflection) lines and count how many are at each reflection point in the pattern.
- Locate any gyration points and figure out the order of rotation at each.
- If there are no gyrations, search carefully for any glide reflections and note the ones that do not cross any reflection lines.
- Finally, if there are translation symmetries (which are common in a repeating pattern) but no other symmetries (no reflections, gyrations, or glide reflections), the pattern must have exactly one wonder-ring, indicated by the signature 0.

3.4 The Magic Theorem

The many wallpaper figures in the previous section illustrate various types of symmetry. To a mathematician, they raise the questions of whether there can be other wallpaper symmetries and how many such two-dimensional symmetries are possible. Figure Ans.12 shows patterns with $*442$ and $*333$ symmetries, but can there be wallpaper, two-dimensional patterns with orbifold signatures of $*452$ and $*345$? Is the number of possible plane symmetries finite, and if so what is it? All theses questions are neatly and systematically answered by the so-called magic theorem of Conway, an elegant mathematical argument which is explained here and is proved in Chapter 6 of reference [Conway 08].

The theorem starts by assigning weights (generally referred to as costs and expressed either in arbitrary units or in Dollars) to the various symbols used in the orbifold signatures. It then claims that the only possible wallpaper symmetries are the 17 combinations of costs that add up to the magic number 2. This puts the limit 17 on the number of plane (two-dimensional) wallpaper symmetry groups and helps determine the orbifold signature of any given pattern. The costs are listed in Table 3.24 and it is easy to verify that all the orbifold signatures illustrated in the previous section have costs of exactly 2 units.

Here are some examples: The total cost of $*632$ is $1+5/6+2/3+1/2=(6+5+2+1)/6=2$. The cost of $2*22$ is $1/2+1+1/4+1/4=2$, and the cost of $**$ is $1+1=2$. In addition to limiting the costs to 2 and the number of wallpaper symmetries to 17,

Symbol	Cost	Symbol	Cost
O	2	$*$ or $\times$	1
2	1/2	2	1/4
3	2/3	3	1/3
4	3/4	4	3/8
5	4/5	5	2/5
6	5/6	6	5/12
⋮	⋮	⋮	⋮
N	$\frac{N-1}{N}$	N	$\frac{N-1}{2N}$

Table 3.24: Costs of Orbifold Symbols.

the magic theorem has another practical application, it helps in identifying, verifying, and sometimes also completing the orbifold signature of a given pattern. A pattern may look rich in symmetries and may trick us to believe we have discovered all kinds of kaleidoscopic lines and gyrations, only to realize later that they do not exist.

Once we have examined a given pattern and decided on its signature, we should verify (1) that the signature's cost is 2, and (2) that it appears in Table 3.24. If the answer to any of these points is negative, we should go back and reexamine the pattern. A typical example is the brick pattern of Figure 3.21. It is easy to locate its reflection axes and start a $*22$ signature, but its low cost $(1+1/4+1/4)$ indicates that a symmetry of cost 1/2 is still missing, thereby providing a hint to look for an order-2 gyration.

We are now ready to explain the unexpected result of 17. This number is simply the number of ways to add entries of Table 3.24 and end up with a total cost of 2. A manual, slightly tedious search of the table produces the following 17 results: We start with just blue entries. No blue entry has a cost of 2. No two entries add up to 2. The only sets of three blue entries with a sum of 2 are 632, 442, and 333. Similarly, the only candidate set of four entries is 2222. Finally, there is the blue wonder-ring 0.

We next notice that the red numeric entries of the table have half the costs of the corresponding blue entries, and that the star's cost is 1. This implies that the all-red combination of costs are similar to the all-blue, i.e., $*632$, $*442$, $*333$, and $*2222$. Further search adds signature $**$ to this set. Finally, there are the hybrid red-blue entries and those that include a $\times$. They are also identified by a simple search and they turn out to be $3*3$, $4*2$, $2*22$, $22\times$, $22*$, $**$, $\times\times$, and $*\times$.

These 17 wallpaper symmetry types, sometimes also called the 17 plane crystallographic groups, are summarized in Table 3.25.

3.5 Orbifold Examples

Figures 3.26 through 3.42 illustrate color examples of the 17 plane symmetry groups. They were generated on the iPad with the beautiful *iOrnament* and *iOrnament Crafter* apps, reference [iOrnament 24], both by Jürgen Richter-Gebert. In addition to the symmetry examples, the figures include variations such as wrapping symmetry patterns around spheres and eggs, and distorting them in limit circles, à la Escher.

Signature	Descritpion
O	just translation
**	parallel mirrors
××	parallel glidelines
*×	parallel mirror and miracles alternating
2222	none
*2222	mutually perpendicular mirrors
2*22	alternating mutual perpendicular mirrors and glidelines
22*	mirrors perpendicular to glidelines
22×	mutually perpendicular glidelines
333	none
*333	all rotation centers located on mirror lines
3*3	an order-3 gyration bounded by three mirrors
442	none
*442	all rotation centers located on mirror lines
4*2	order-4 gyration bounded by a rectangle of mirror lines
632	none
*632	reflection lines form a triangle

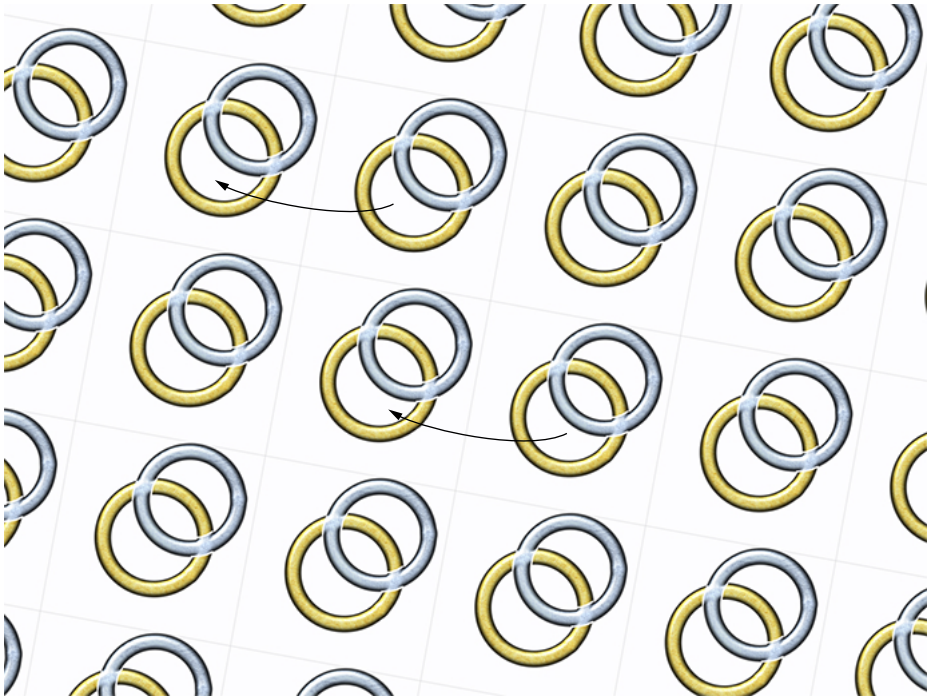
Table 3.25: The 17 Wallpaper Groups.

3.6 Two-Dimensional Transformations

The vast field of computer graphics deals with computer-generated objects. Such objects are made of synthetic surfaces, they are manipulated in the computer and eventually rendered to look real. An important part of manipulating objects is to transform them. The obvious example is rotation. In real life, we see objects in three dimensions, because we see them with both our eyes. In the computer, we see the same three-dimensional objects on a flat screen or on paper, where they look flat and two-dimensional. What we see is a two-dimensional projection of a three-dimensional object, and such a projection often obscures important object details. Thus, in order to see a three-dimensional object in its entirety on a flat output device, it has to be rotated and viewed from different directions.

Rotation is one of the important geometric transformations, a set which also includes translation, reflection, and shearing (Figure 3.43). There are also less important transformations, such as glide reflection, but this discussion concentrates on the symmetry aspects of the transformations, and especially on the relations between rotations and reflections. We start with a look at the linear mathematics underlying the two-dimensional transformations.

We use the notation $\mathbf{P} = (x, y)$ for a point and $\mathbf{P}^* = (x^*, y^*)$ for the transformed point. We are looking for a simple, fast transformation rule, so it is natural to try a linear transformation, i.e., a mathematical rule that functions of x, but not x^2 or anything more complex. The simplest linear transformation is $x^* = ax + cy$ and $y^* = bx + dy$,



Translation only

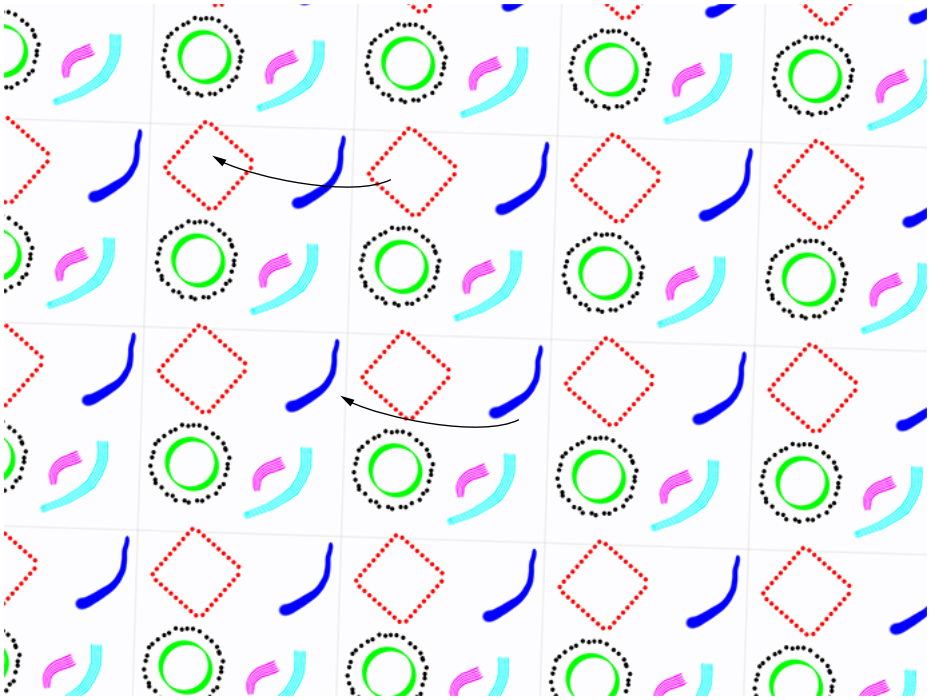


Figure 3.26: Symmetry Group P1, Orbifold O.

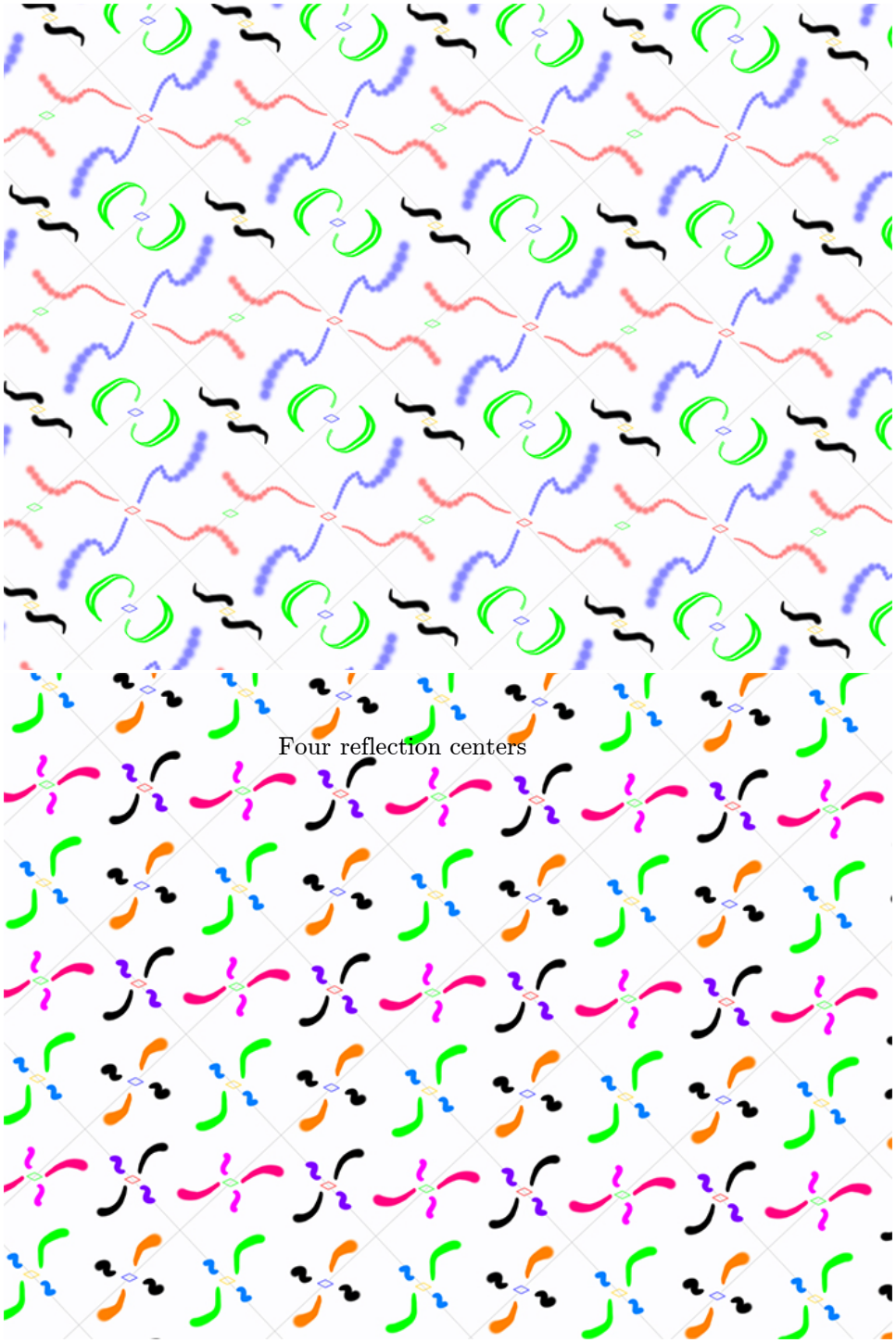


Figure 3.27: Symmetry Group P2, Orbifold 2222.

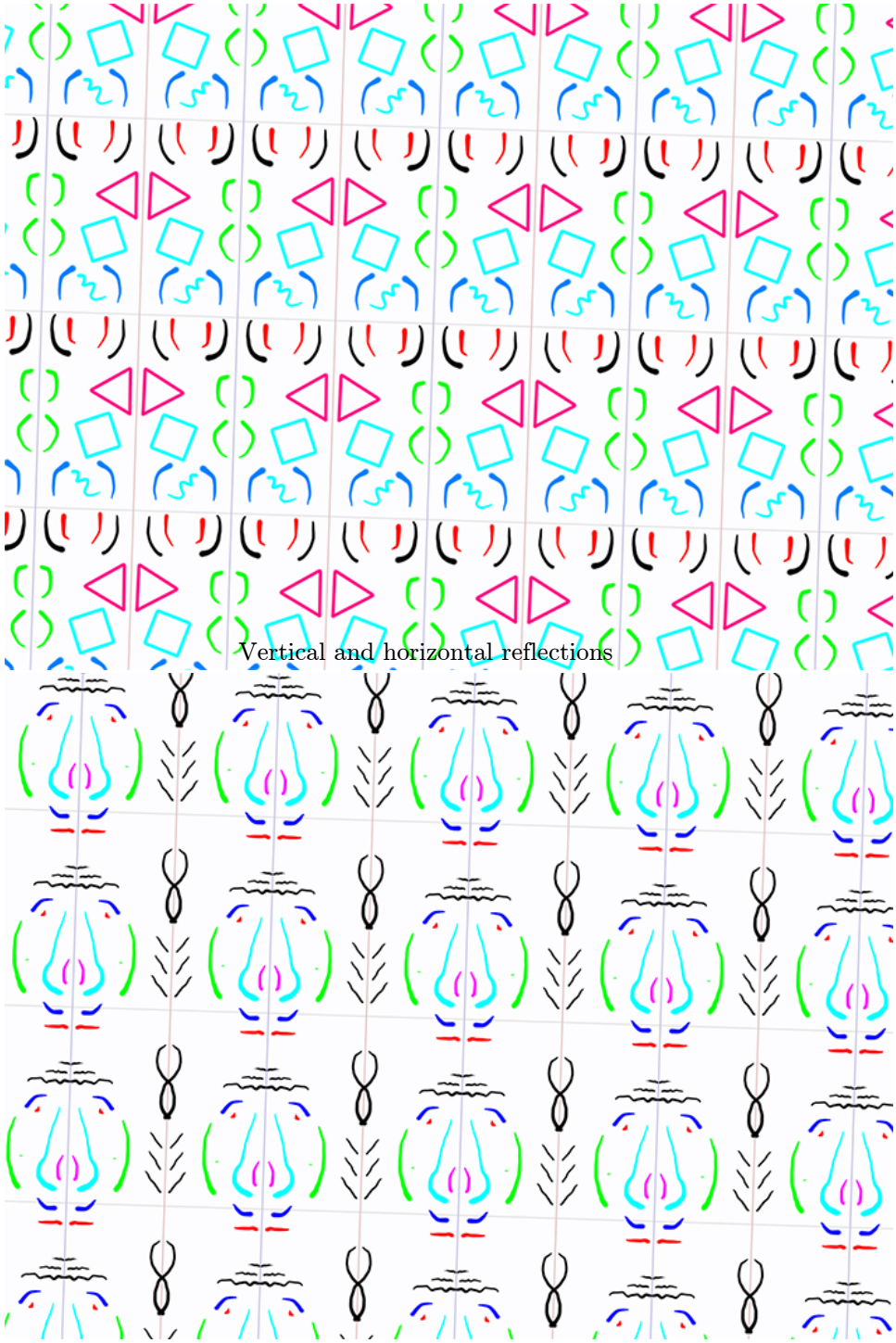


Figure 3.28: Symmetry Group Pm, Orbifold **.

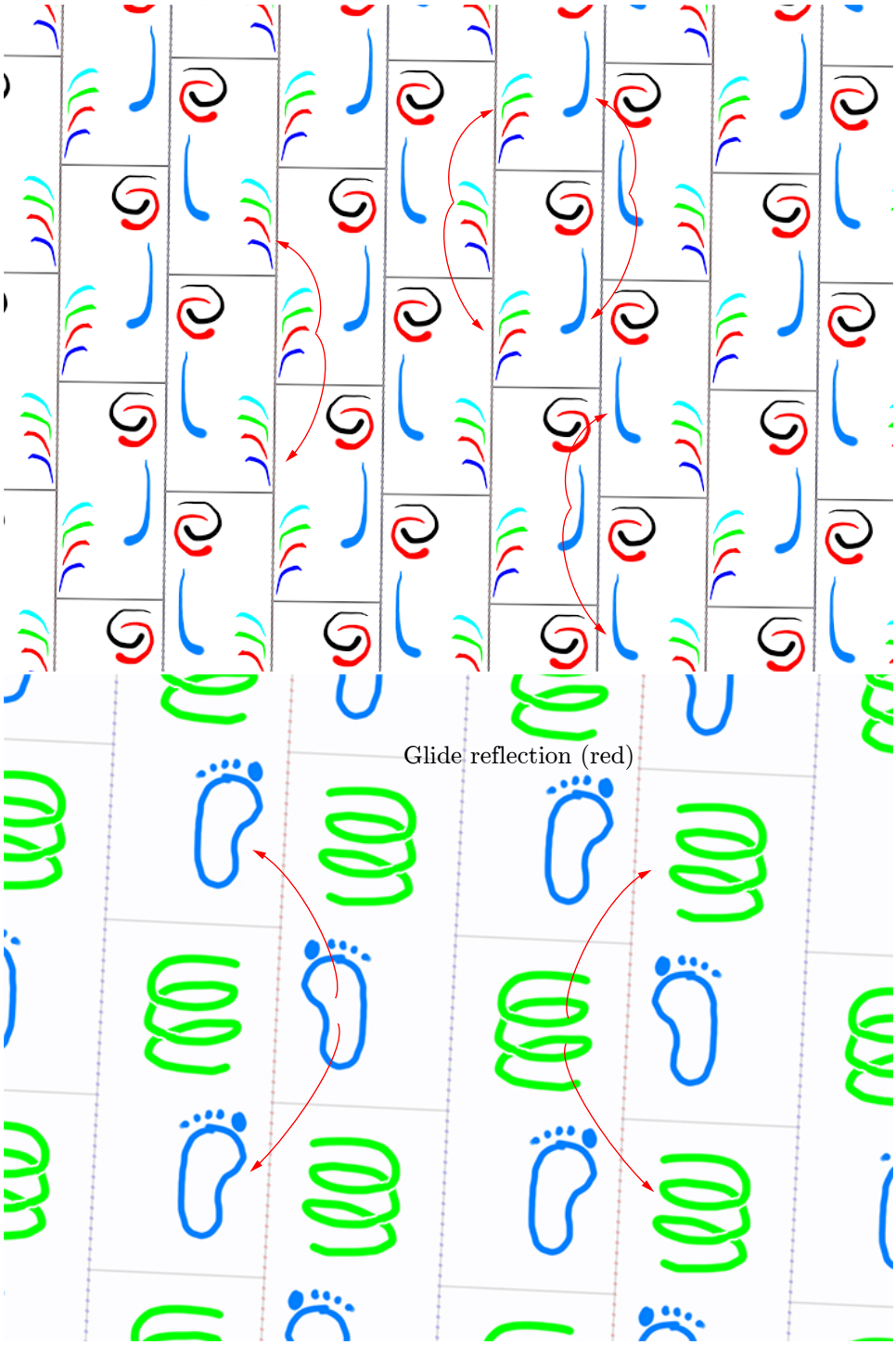


Figure 3.29: Symmetry Group Pg, Orbifold xx.

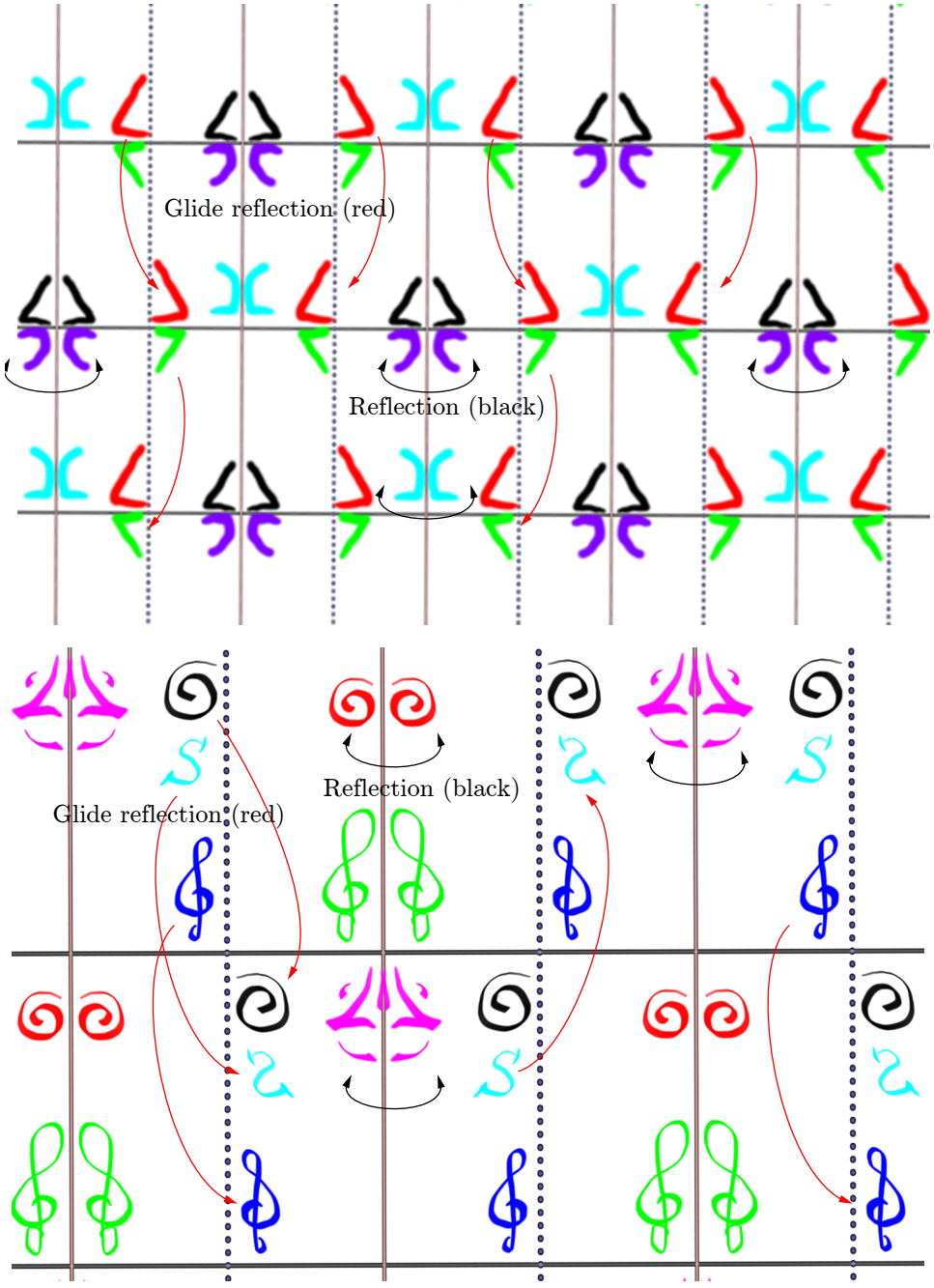


Figure 3.30: Symmetry Group Cm, Orbifold *x.

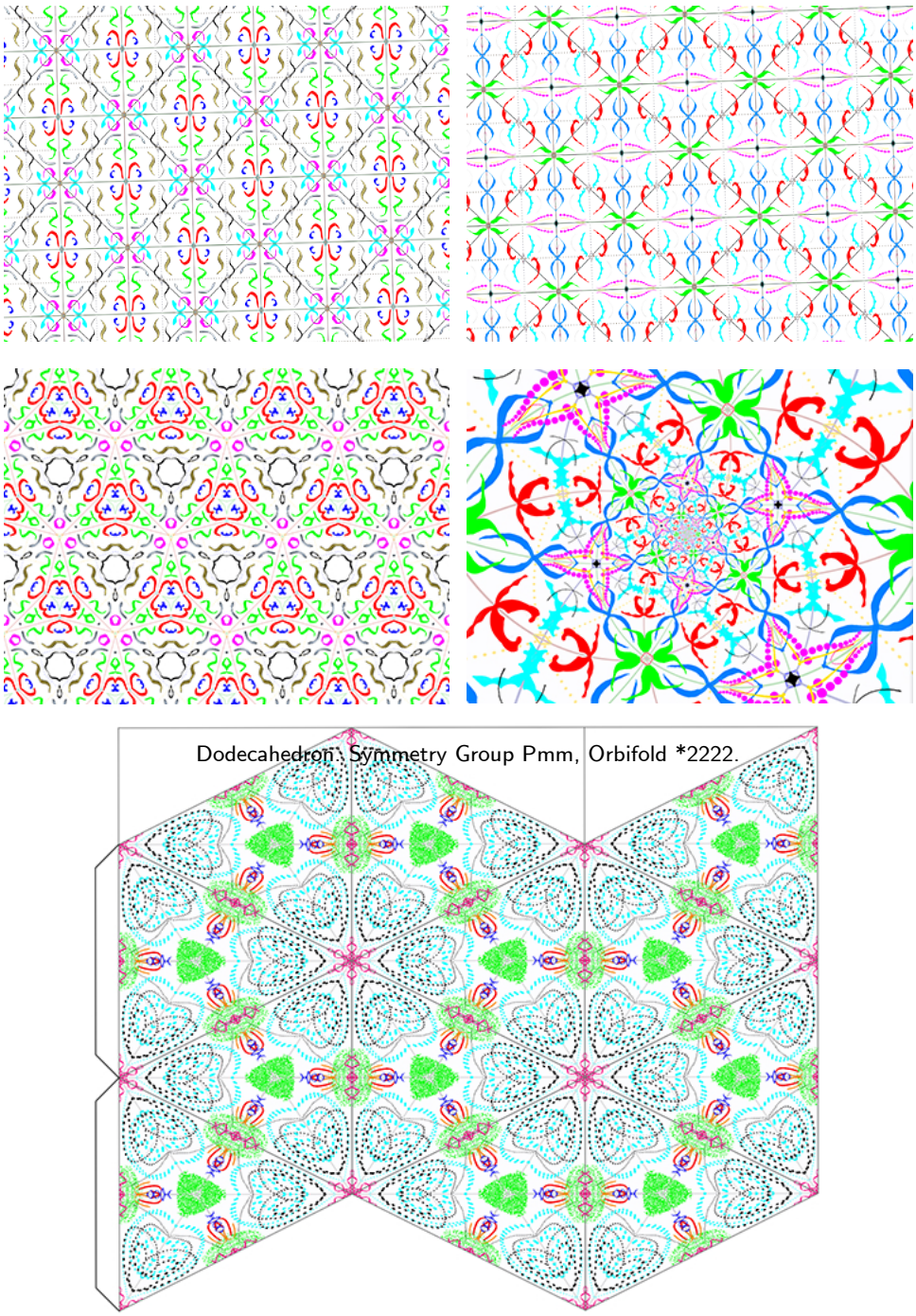


Figure 3.31: Symmetry Group Cmm, Orbifold 2*22.

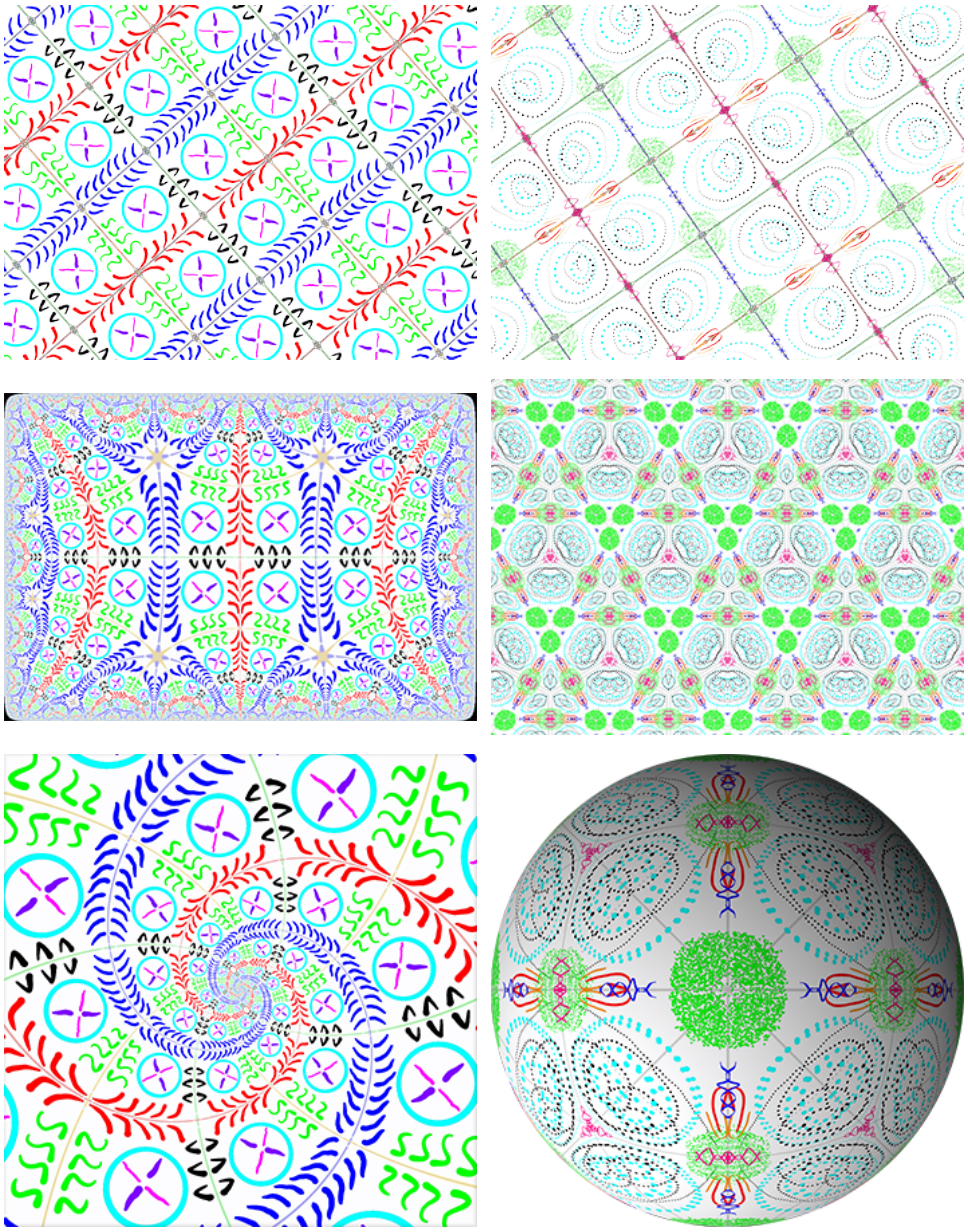


Figure 3.32: Symmetry Group Pmm, Orbifold *2222.

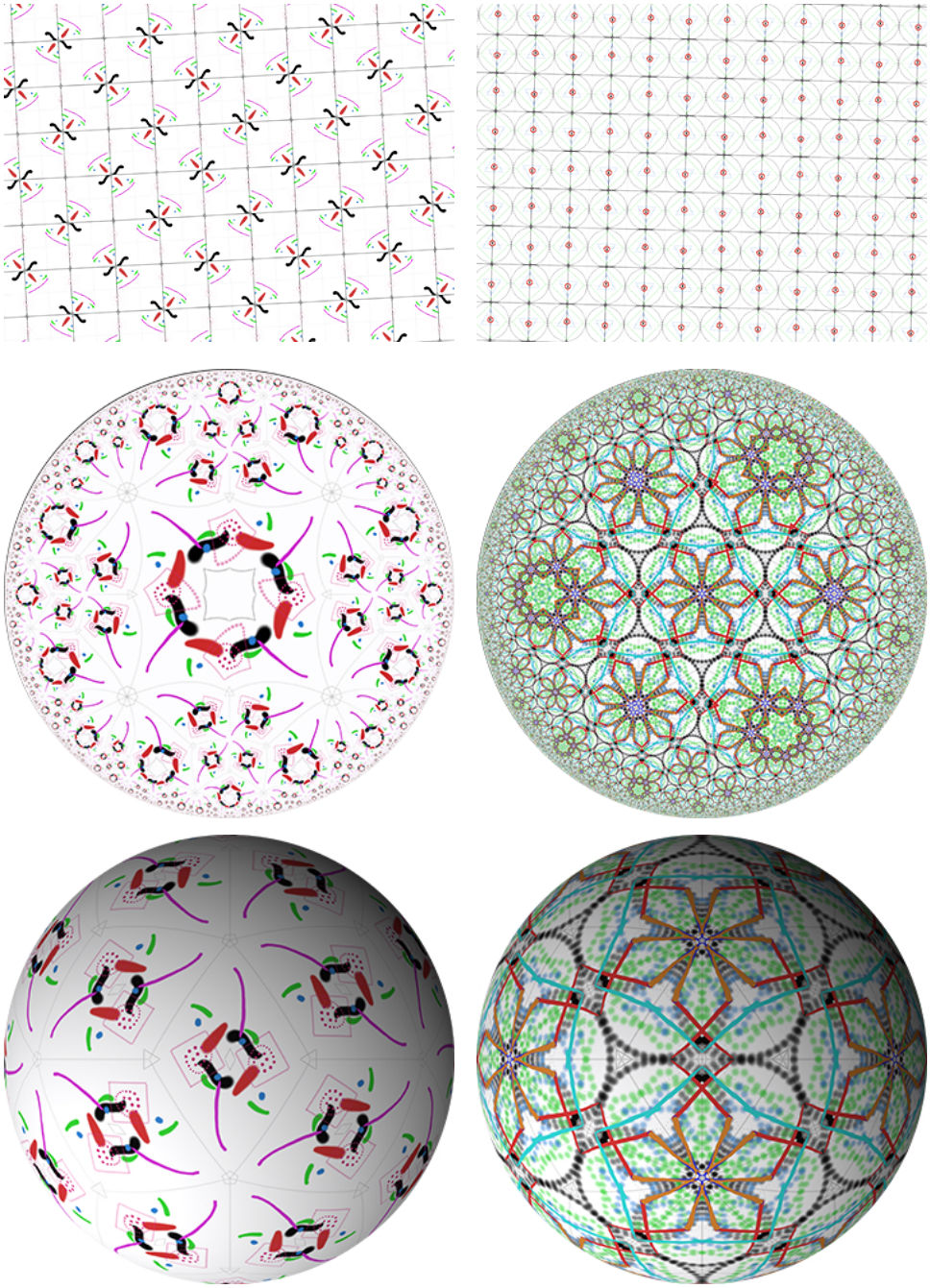


Figure 3.33: Symmetry Group Pgg, Orbifold 22*.

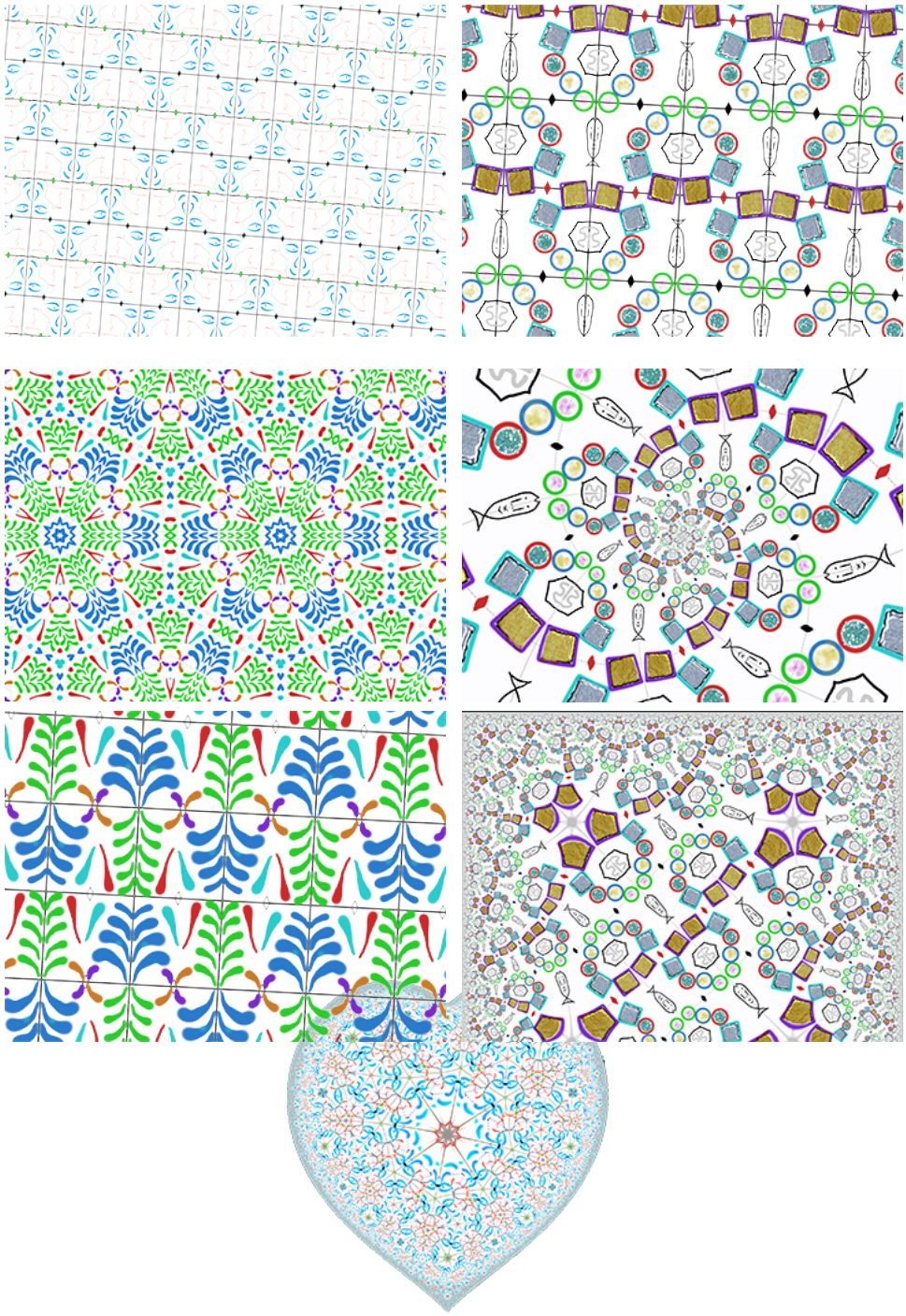


Figure 3.34: Symmetry Group Pgg, Orbifold 22x.

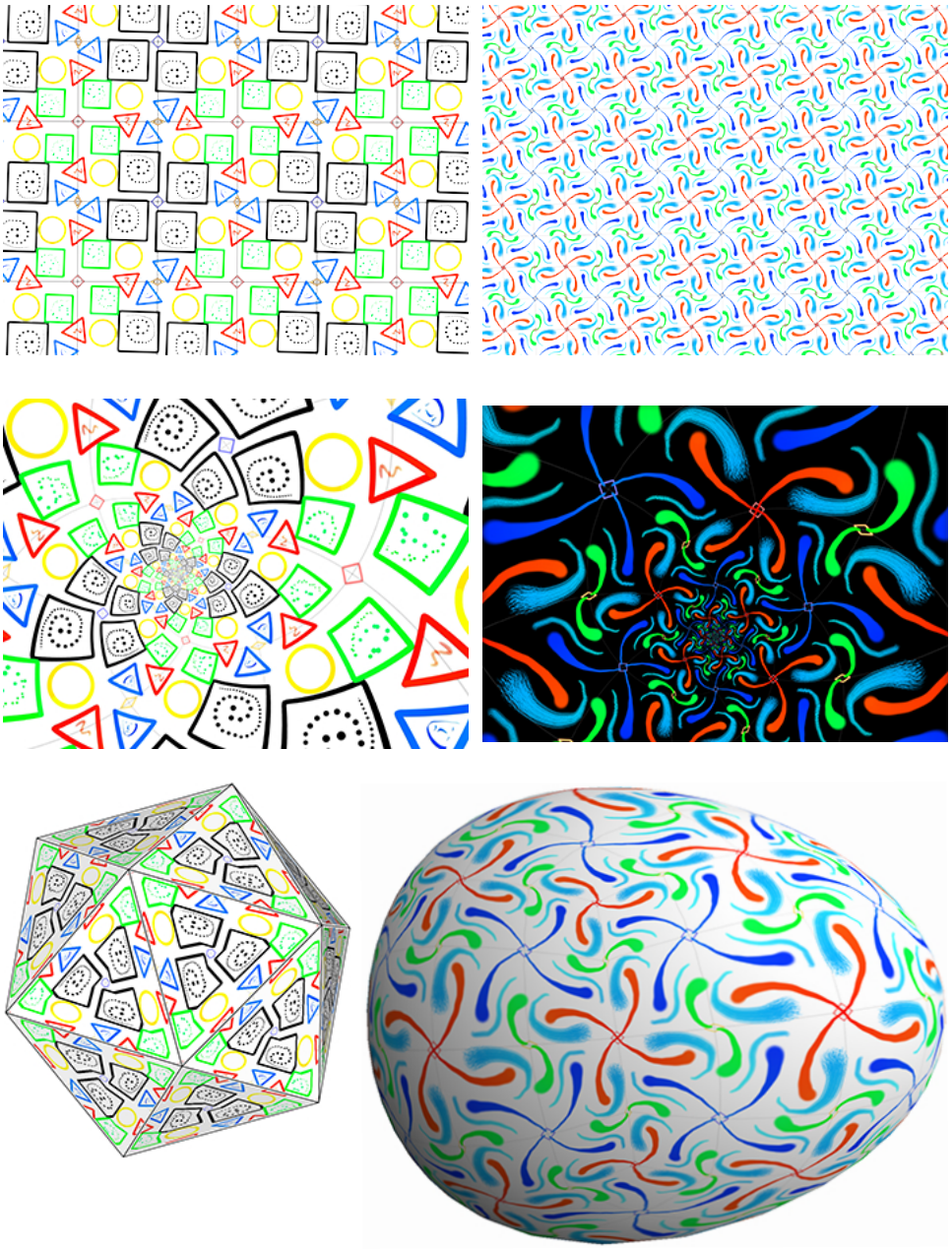
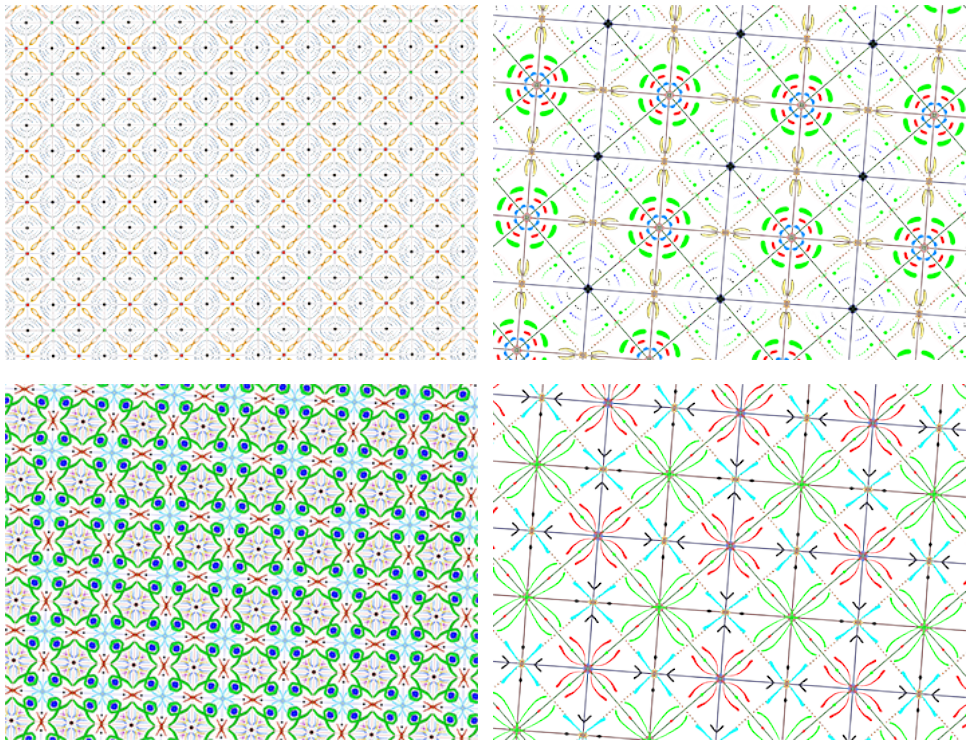


Figure 3.35: Symmetry Group P4, Orbifold 442.



Symmetry Group P4m, Orbifold *442.

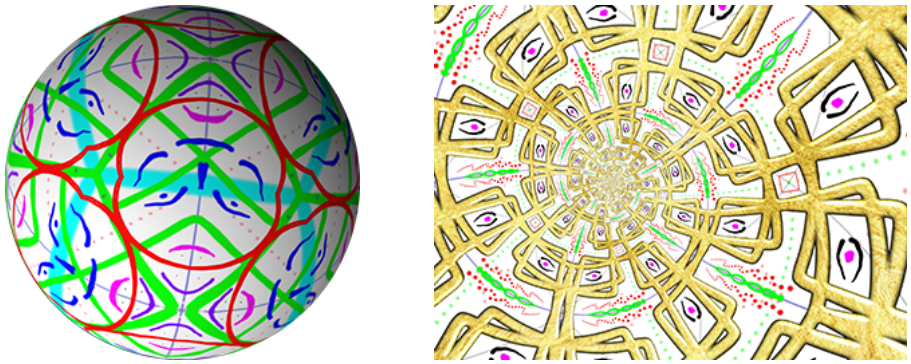


Figure 3.36: Symmetry Group P4g, Orbifold 4*2.

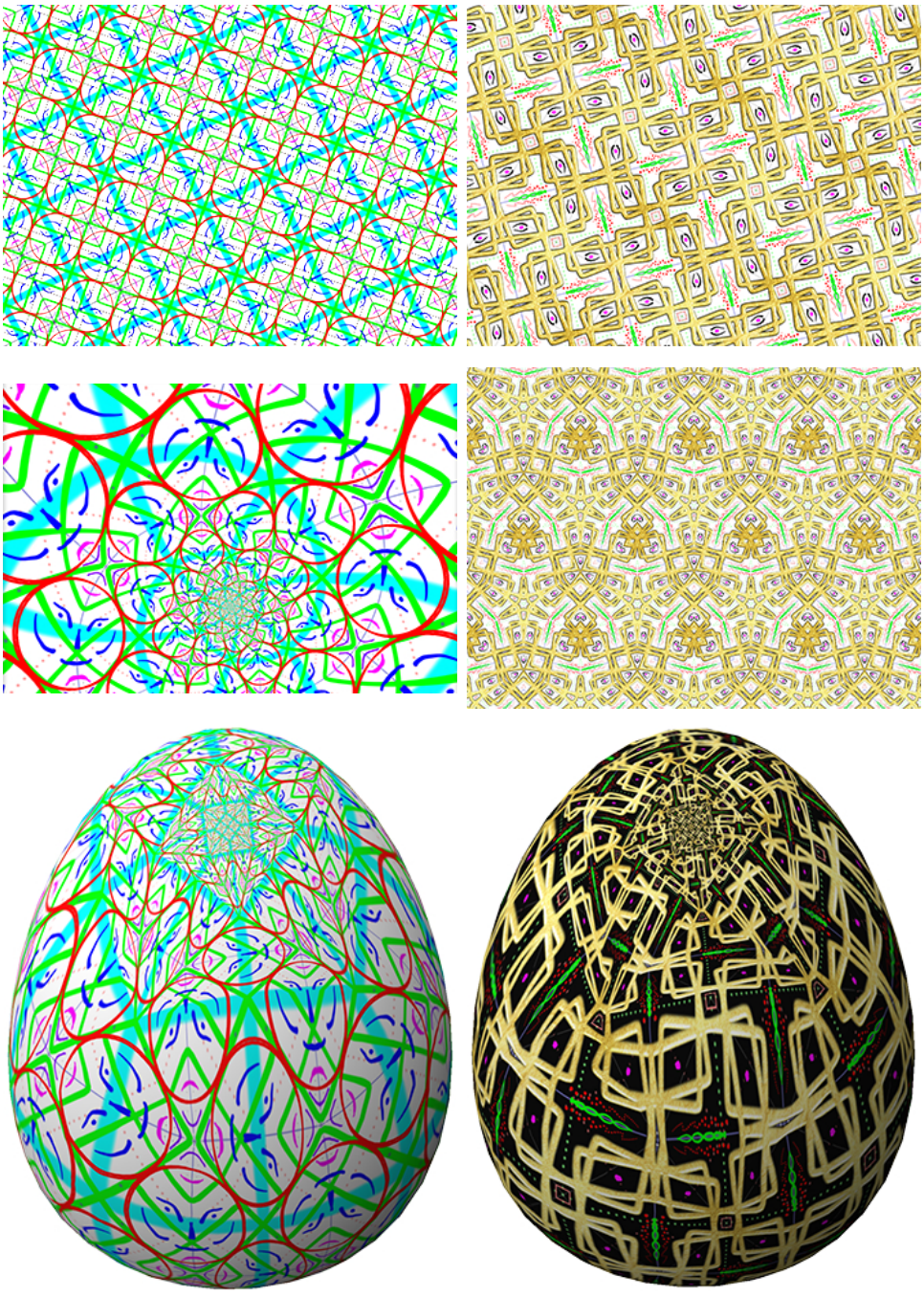


Figure 3.37: Symmetry Group P4g, Orbifold 4*2.

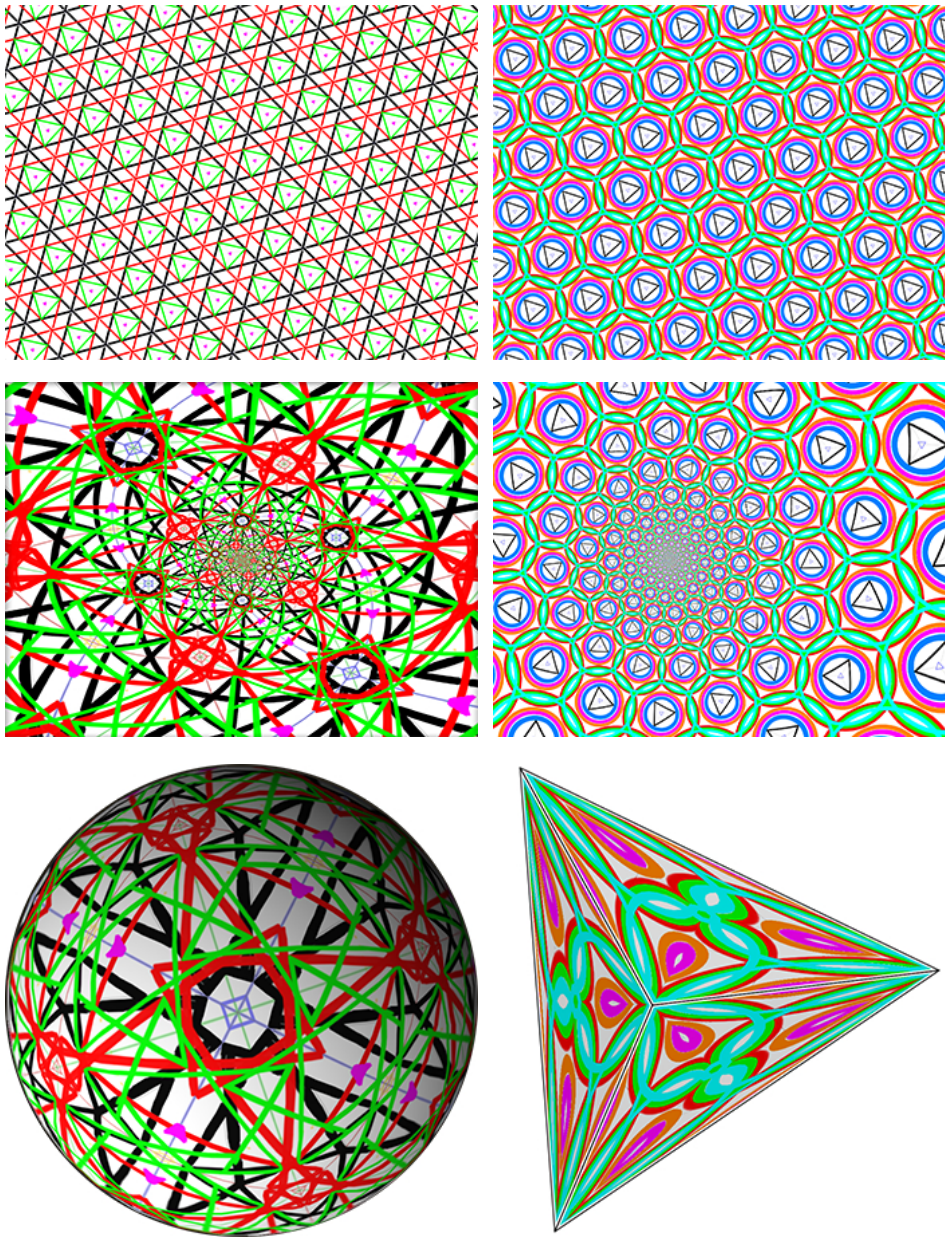


Figure 3.38: Symmetry Group P3, Orbifold 333.

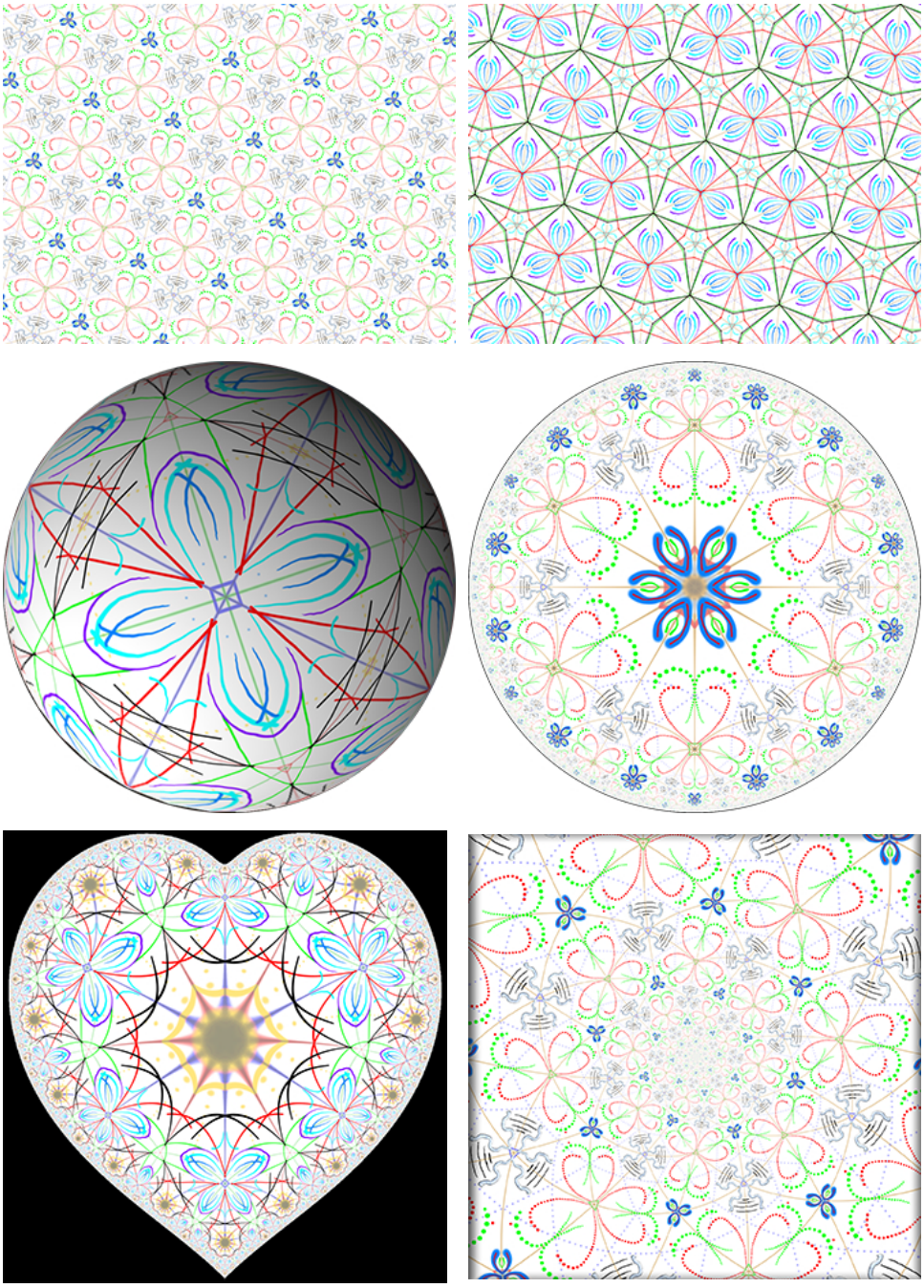


Figure 3.39: Symmetry Group P3m1, Orbifold *333p.



Figure 3.40: Symmetry Group P31m, Orbifold 3*3.

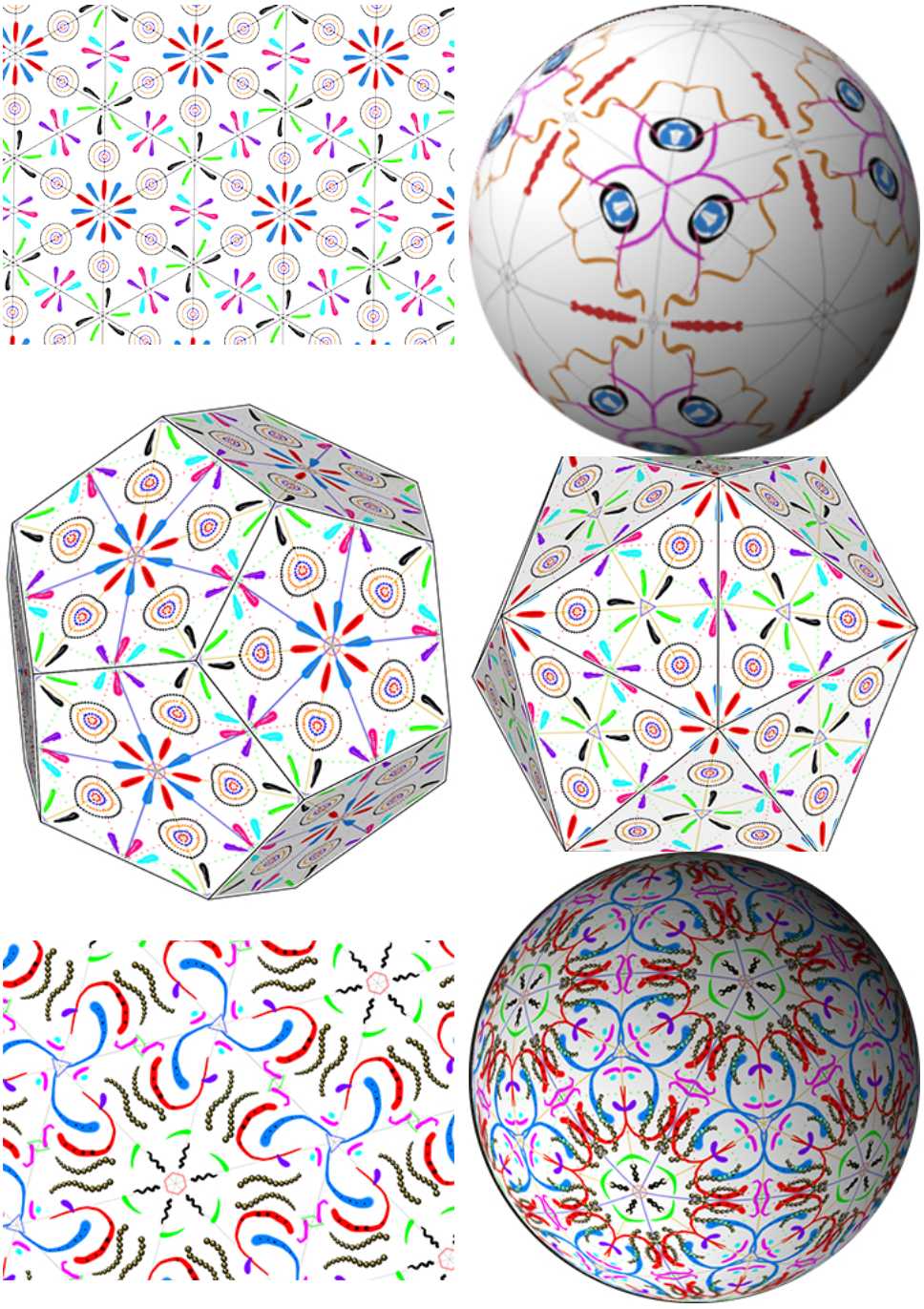


Figure 3.41: Symmetry Group P6, Orbifold 632.

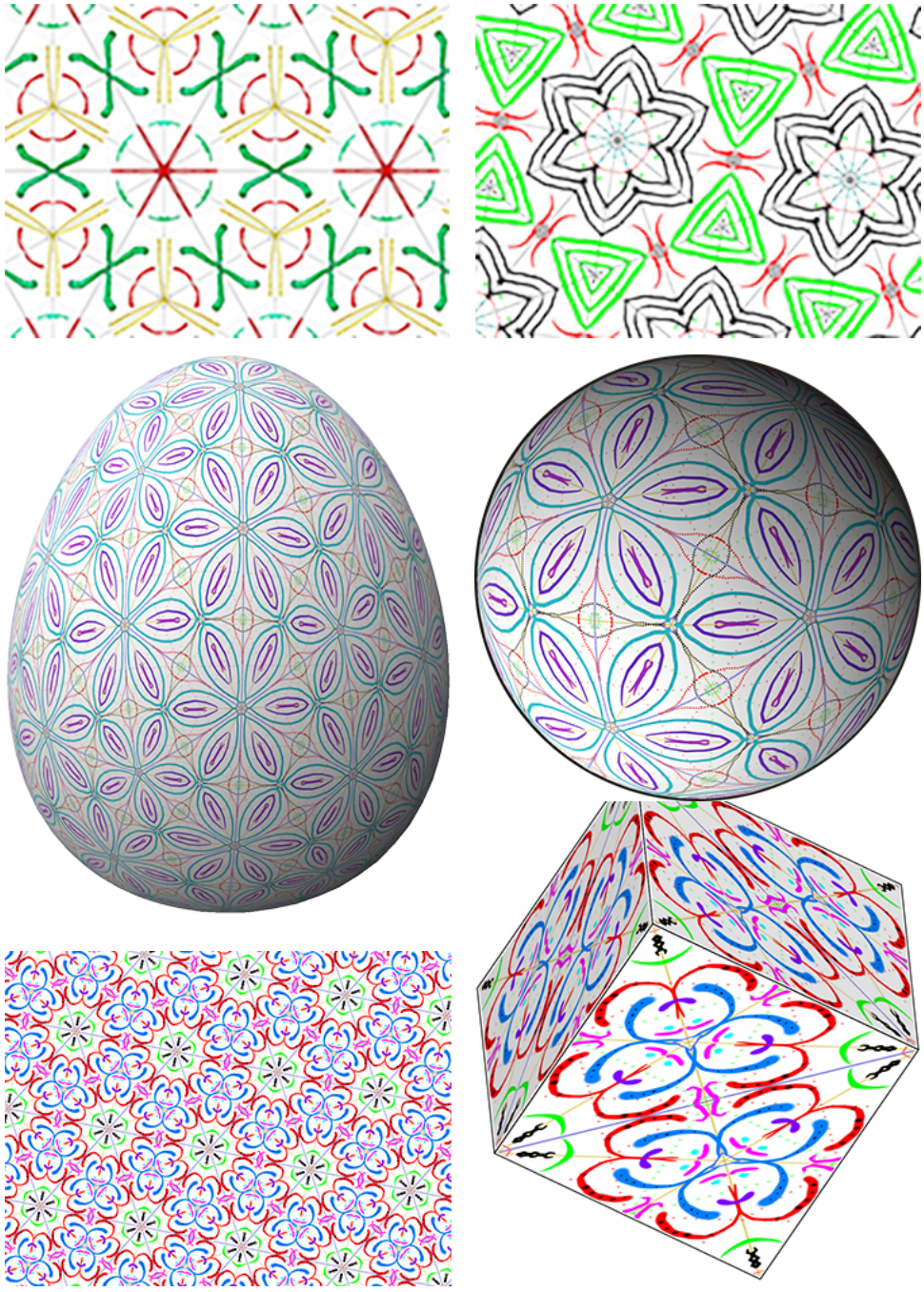


Figure 3.42: Symmetry Group P6m, Orbifold *632.



Original Y reflected Sheared Y scaled Rotated

Figure 3.43: Two-Dimensional Transformations.

in which each of the new coordinates is a linear combination of the two old ones. This transformation can be written $\mathbf{P}^* = \mathbf{PT}$, where $\mathbf{T}$ is the 2×2 matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$.

To understand the functions of the four matrix elements, we start by setting $b = c = 0$. The transformation becomes $x^* = ax$, $y^* = dy$. Such a transformation is called *scaling*. If applied to all the points of an object, all the x dimensions are scaled by a factor of a and all the y dimensions are scaled by a factor of d. Note that a and d can also be less than 1, causing shrinking of the object. If any of a and d equal -1, the transformation is a *reflection*. Any other negative values cause both scaling and reflection.

Note that scaling an object by factors of a and d changes its area by a factor of $a \times d$, and that this factor is also the value of the determinant of the scaling matrix $\left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$.

Here are examples of scaling and reflection. In $\mathbf{A}$, the y coordinates are scaled by a factor of 2. In $\mathbf{B}$, the x coordinates are reflected. In $\mathbf{C}$, the x dimensions are shrunk to 0.001 their original values. In $\mathbf{D}$, the figure is shrunk to a vertical line.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}; \quad \mathbf{B} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}; \quad \mathbf{C} = \begin{pmatrix} 0.001 & 0 \\ 0 & 1 \end{pmatrix}; \quad \mathbf{D} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

◇ **Exercise 3.16:** What scaling transformation changes a circle to an ellipse?

The next step is to set $a = 1$ and $d = 1$ (no scaling or reflection) and explore the effect of b and c. The transformation becomes $x^* = x + cy$, $y^* = bx + y$. We first try the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and transform the four points at $(1,0)$, $(3,0)$, $(1,1)$, and $(3,1)$. They are transformed to $(1,1)$, $(3,3)$, $(1,2)$, and $(3,4)$. When we plot the original and the transformed points (Figure 3.44a) it becomes obvious that the original rectangle has been sheared vertically and was transformed into a parallelogram. A similar effect occurs when we try the matrix $\left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. The quantities b and c are therefore responsible for *shearing*. Figure 3.44b shows the connection between shearing and the operation of scissors. This is the reason for the name shearing.

◇ **Exercise 3.17:** Apply the shearing transformation $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ to the four points $(1,0)$, $(3,0)$, $(1,1)$ and $(3,1)$. What are the transformed points? What geometric figure do they represent?

The next important transformation is rotation. In Euclidean geometry, a rotation is an example of an isometry, a transformation that moves points without changing

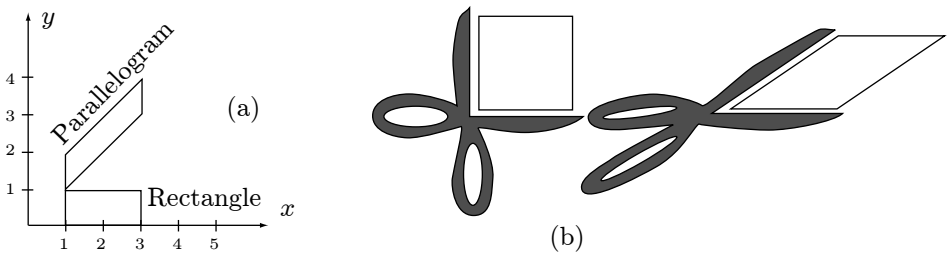


Figure 3.44: Scissors and Shearing.

the distances between them. Rotations are distinguished from other isometries by two additional properties: they leave (at least) one point fixed, and they leave "handedness" unchanged. In contrast, a translation moves every point, a reflection exchanges left- and right-handed ordering, a glide reflection does both, and an improper rotation combines a change in handedness with a normal rotation.

Figure 3.45 shows a point $\mathbf{P}$ rotated clockwise about the origin through an angle θ to become $\mathbf{P}^*$. Simple trigonometry gives $x = R\cos\alpha$ and $y = R\sin\alpha$. From this, we get the expressions for x^* and y^*

$$x^* = R\cos(\alpha-\theta) = R\cos\alpha\cos\theta + R\sin\alpha\sin\theta = x\cos\theta + y\sin\theta,$$
$$y^* = R\sin(\alpha-\theta) = -R\cos\alpha\sin\theta + R\sin\alpha\cos\theta = -x\sin\theta + y\cos\theta.$$

Hence, the clockwise rotation matrix in two dimensions is

$$\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}, \quad \begin{matrix}\text{which also}\\ \text{equals}\\ \text{the product}\end{matrix} \quad \begin{pmatrix}\cos\theta & 0\\ 0 & \cos\theta\end{pmatrix}\begin{pmatrix}1 & -\tan\theta\\ \tan\theta & 1\end{pmatrix}. \tag{3.2}$$

This shows that any rotation in two dimensions is a combination of scaling (and, perhaps, reflection) and shearing; a neat result that's true for all angles satisfying $\tan\theta \neq \infty$.

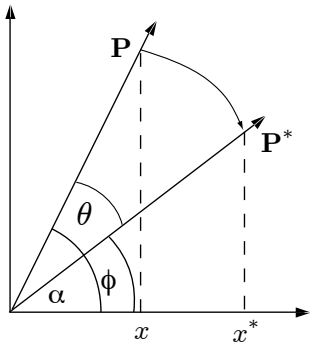


Figure 3.45: Clockwise Rotation.

◇ **Exercise 3.18:** Show how a 45° rotation can be achieved by scaling followed by shearing.

◇ **Exercise 3.19:** Derive the rotation matrix in two dimensions by using the polar coordinates (r, θ) of points instead of the Cartesian coordinates (x, y).

Matrix $\mathbf{T}_1$ below rotates counterclockwise. Matrix $\mathbf{T}_2$ reflects about the line $y = x$, and matrix $\mathbf{T}_3$ reflects about the line $y = -x$. Note the determinants of these matrices. In general, a determinant of $+1$ indicates pure rotation, whereas a determinant of -1 indicates pure reflection.

$$\mathbf{T}_1 = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}; \qquad \mathbf{T}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \qquad \mathbf{T}_3 = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}. \tag{3.3}$$

◇ **Exercise 3.20:** Show that a y reflection (i.e., reflection about the x axis) followed by a reflection through the line $y = -x$ produces pure rotation.

◇ **Exercise 3.21:** Show that the transformation matrix

$$\begin{pmatrix} \frac{1-t^2}{1+t^2} & \frac{2t}{1+t^2} \\ \frac{-2t}{1+t^2} & \frac{1-t^2}{1+t^2} \end{pmatrix}$$

produces pure rotation.

◇ **Exercise 3.22:** For what values of A does the following matrix represent pure rotation and for what values does it represent pure reflection?

$$\begin{pmatrix} a/A & b/A \\ -b/A & a/A \end{pmatrix}.$$

We conclude with several surprising results showing unexpected connections between linear transformations.

■ Two consecutive reflections about axes that intersect are the same as a rotation about the point of intersection. A rotation leaves one point unchanged, while a reflection is fully determined by a line of reflection. Two reflections are therefore determined by two lines, and two lines generally intersect at a point. This point is the center of the rotation that is equivalent to the reflections. The object of Figure 3.46 is first reflected about the red line and then about the green line, ending up at point 3. Rotating it from point 1 about the intersection point (in black) of the two lines, would also have landed it at the same point.

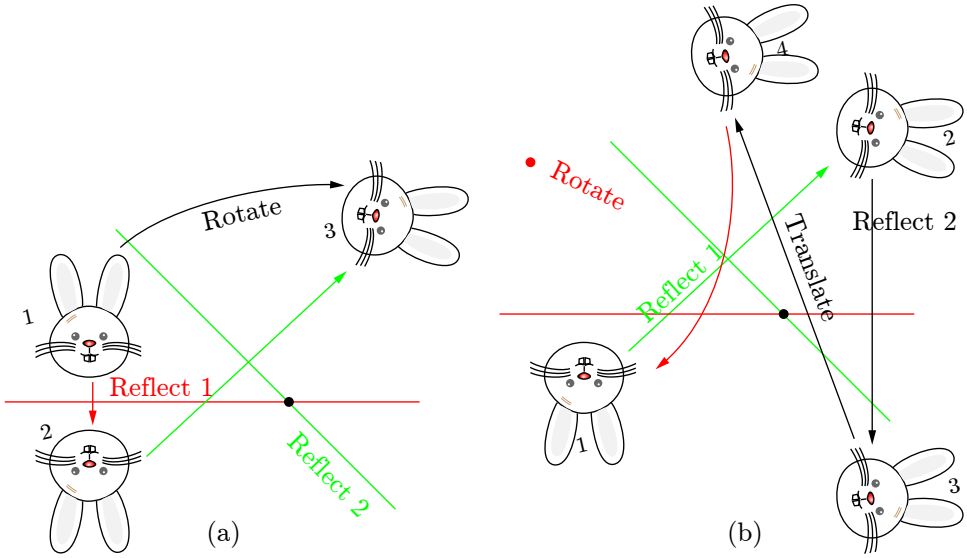


Figure 3.46: Two Reflections and a Rotation.

◇ **Exercise 3.23:** Discuss the cases where the two reflection lines are either parallel or identical.

■ A rotation followed by a translation is always equivalent to another rotation. Figure 3.46b illustrates an example. Object 1 is reflected through the green line to 2, then reflected again about the red line to 3. These two transformations amount to a rotation, and then the object is translated to an arbitrary location 4. This combination of a rotation followed by a translation is then negated by computing a new center of rotation (the red dot) and rotating the object about this center from 4 back to its original location 1.

Here is how to compute the new center of rotation. The mathematics of two-dimensional rotations is simple. Given a two-dimensional point $\mathbf{x} = (x, y)$ and a rotation angle θ, the new location (x', y') of the point after a rotation by θ about the origin is given by the matrix product

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = R\,\mathbf{x},$$

Where R is the rotation matrix for angle θ and the notation $\mathbf{x}$ in boldface implies a point or a vector. The mathematics of two-dimensional translations is even simpler. Translating a point $\mathbf{x}$ by a vector $\mathbf{b}$ is denoted by the operator $T(\mathbf{x}) = \mathbf{x} + \mathbf{b}$.

A rotation followed by a translation is the product $T \cdot R$, and in order for this to be a rotation, there must be a unique point $\mathbf{p}$ in space that is left fixed after this combination of operators. Thus, $(T \cdot R)(\mathbf{p}) = \mathbf{p}$ or $R\,\mathbf{p} + \mathbf{b} = \mathbf{p}$. The solution is $(I - R)\mathbf{p} \stackrel{\text{def}}{=} A\mathbf{p} = \mathbf{b}$ (where I is the identity matrix) or $\mathbf{p} = A^{-1}\mathbf{b}$. Such a solution exists because the determinant of A is nonzero. Thus, if $\mathbf{x}$ and θ are known, matrix A

can be computed and inverted to obtain $\mathbf{p}$. Figure 3.47 lists Mathematica code for this computation.

◇ **Exercise 3.24:** Show why the determinant of A is nonzero.

```
t = 45.; v = {5, 9};(*choose rotation angle and point*)
R = {{Cos[t Degree], -Sin[t Degree]}, {Sin[t Degree], Cos[t Degree]}};
A = IdentityMatrix[2] - R;
Det[A]; (*verify nonzero*)
b = {2, 3}; (*choose translation vector*)
p = LinearSolve[A, b] (*p = new center of rotation*)
```

Figure 3.47: Mathematica Code for Computing Point $\mathbf{p}$.

■ Rotation and reflection matrices are close relatives. The proof of this unexpected result depends on the fact that our 2×2 transformation matrices are orthogonal, so we start with a short discussion of this concept.

An orthogonal matrix is a square matrix whose rows are orthogonal and of length 1, and similarly for its columns. Because of the length 1 requirement, such a matrix is also referred to as orthonormal. Examples are:

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad \frac{1}{3}\begin{pmatrix} 2 & -2 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & -2 \end{pmatrix}.$$

It's easy to see that, e.g., the length of row $(1/3)(2,-2,1)$ is $\sqrt{2^2+(-2)^2+1^2}/3 = 1$, and that the dot product $(1,2,2)\cdot(2,1,-2) = 2+2-4$ is zero. The high-level Wolfram Language offers the command `OrthogonalMatrixQ[m]` to test a matrix for orthogonality.

Given the 2×2 orthogonal matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, the orthogonality definition implies the three constraints $a^2+b^2 = 1$, $c^2+d^2 = 1$, and $ac+db = 0$. These in turn mean that the four matrix elements must be numbers in the interval $[-1,1]$, and can therefore be considered the sines or cosines of certain angles. The constraints above imply that if $a = \cos(\theta)$ and $b = \sin(\theta)$ for some angle θ, then either $c = -b$ and $d = a$, or $c = b$ and $d = -a$, which is why the above matrix has one of the forms

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

The first of these describes a rotation by θ about the origin, and the second is a reflection about the line $y = (\theta/2)x$.

We can therefore conclude that all 2×2 orthogonal matrices are rotation or reflection

matrices; a nonintuitive result that yields the following identities:

$$\begin{aligned}
\text{Rot}(\theta)\text{Rot}(\phi) &= \text{Rot}(\theta+\phi),\\
\text{Ref}(\theta)\text{Ref}(\phi) &= \text{Rot}(2\theta-2\phi),\\
\text{Rot}(\theta)\text{Ref}(\phi) &= \text{Ref}(\phi+\frac{1}{2}\theta),\\
\text{Ref}(\phi)\text{Rot}(\theta) &= \text{Ref}(\phi-\frac{1}{2}\theta).
\end{aligned}$$

- The set of all reflections in lines through the origin and rotations about the origin, together with the operation of composition of reflections and rotations, forms a group. The group has an identity: Rot(0). Every rotation Rot(r) has an inverse Rot(-r). Every reflection Ref(f) is its own inverse. Composition has closure and is associative, since matrix multiplication is associative.

Perpendicular, Normal, and Orthogonal

The terms above are commonly used in mathematical texts. They all have to do with straight angles, but their use in practice is somewhat different.

- The term perpendicular is used for two linear objects. Two straight lines at 90° are perpendicular, as are also a line and a flat plane and two planes, if their intersections form right angles. Perpendicularity is a special case of orthogonality.

- Normal is the term used when a line is at right angle to a point on a curve or on a surface. More accurately, the normal to a curve at a point **P** is the direction perpendicular to the tangent of the curve at **P**. Similarly, if a surface is smooth at a point, it has a tangent plane at the point, and the direction perpendicular to this plane is the normal of the surface at the point.

- Orthogonal is a more general term, and is used in science, engineering, art, and computers in addition to mathematics. In mathematics, orthogonality is the generalization of the term perpendicularity. It is defined for vectors, functions, and polynomials.

Two vectors are orthogonal if their inner product (or dot or scalar product, Page 472) is zero. Thus, $(a,0)\bullet(0,b)$ is a simple example of orthogonal vectors.

An orthogonal matrix is a square matrix whose rows are orthogonal, and similarly for its columns.

A family of functions is orthogonal over an interval $[a,b]$ if for any pair of different functions in the family, the definite integral

$$\int_a^b \bar{f}(x)g(x)\,dx$$

is zero.

A family of polynomials is similarly orthogonal over an interval $[a,b]$ if for any pair of different polynomials in the family, the definite integral

$$\int_a^b \bar{p}(x)q(x)\,dx$$

is zero.

3.7 Symmetry in Tiling

A tessellation or tiling of the plane is the covering of the plane with geometric shapes, referred to as tiles, with no overlaps and no gaps. In mathematics, tessellation can be generalized to higher dimensions.

The mathematical concept of tiling is geometric. In two dimensions, tiling is a partition of the infinite plane into closed sets which are referred to as tiles, without gaps or overlaps. In practice, most floor and wall tiles are periodic. An infinite tiled plane is periodic if it can be translated onto itself in two perpendicular directions. Periodic tiling consists of primitive cells that repeat in two independent directions (Figure reffg:primitv.cell). Each set may be composed of any number of smaller tiles. Nonperiodic tilings exist and can have artistic and practical properties. A nonperiodic tiling is one that cannot be constructed from a single primitive cell.

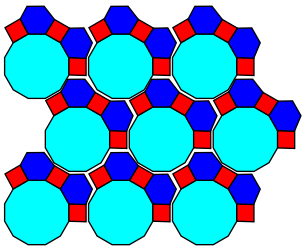


Figure 3.48: A Primitive Cell of Tiles.

Aperiodic tiling is a special case of nonperiodic tiling where small regions may be periodic, but it is always possible to find large regions or patches that are not periodic. Such tiling does not contain arbitrarily large periodic regions.

A huge variety of tiles is available, both for floors and walls. Our interest in tiles is in their application to symmetry. Certain types of tiles, especially ceramic, are made in different shapes and can be laid in configurations that feature various symmetries. Other tiles, such as vinyl, feature patterns and ornaments on their surface, and can also be arranged in symmetric configurations.

We start with short definitions. A polygon is a two-dimensional enclosed figure made by joining three or more straight segments at points that are referred to as vertices. A regular polygon is one whose sides and interior angles are equal. Figure 3.49a illustrates a few such polygons and interior and exterior angles. In an n-sided polygon, the sum of the n interior angles is $(n-2)\cdot 180^\circ$, and the sum of the n exterior angles (blue in the figure) is 360°.

The simplest way to tile a plane is regular tiling, where congruent regular polygons (i.e., having the same shape and size or being mirror images) are laid edge to edge.

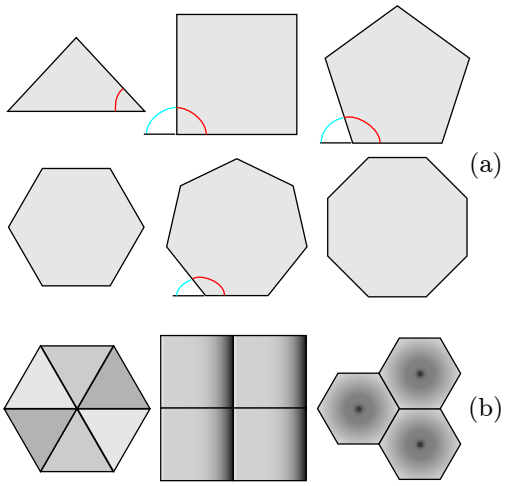


Figure 3.49: Regular Polygons and Angles.

There are only three possible regular tilings (Figure 3.49b), namely triangles (where six equilateral triangles meet at a vertex), squares (where four squares meet at a vertex), and regular hexagons (where three meet at each vertex).

When we relax the condition of congruent regular polygons and allow tiling of different polygons laid edge to edge, the result is eight different types of semiregular tilings, also known as Archimedean or uniform tiling (Figure 3.50).

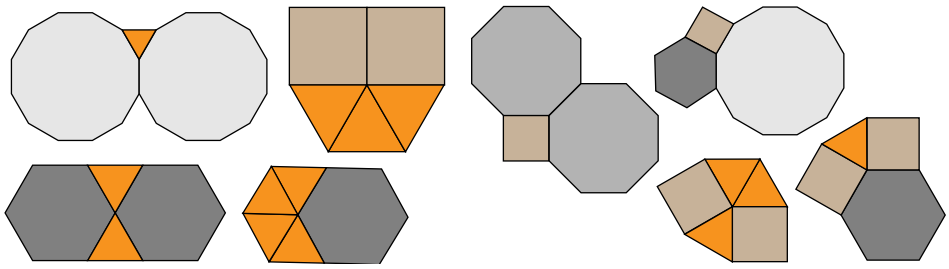


Figure 3.50: Semiregular Tiling of the Plane.

If we relax the requirements that vertices meet at a point and adjacent edges have the same length, we unleash a vast number of ways to tile the plane, some more pleasing to the eye than others (Figure 3.51).

Aperiodic tilings are complex, mathematically interesting, but are considered much less beautiful and pleasing to the eye than the various periodic tiling schemes. Currently (in late 2023), the most important aperiodic tilings are Penrose tiling and Einstein tiling. Penrose tiling, discovered by Roger Penrose in the 1970's, have both reflection symmetry and fivefold rotational symmetry, but not translational symmetry. It is possible to find

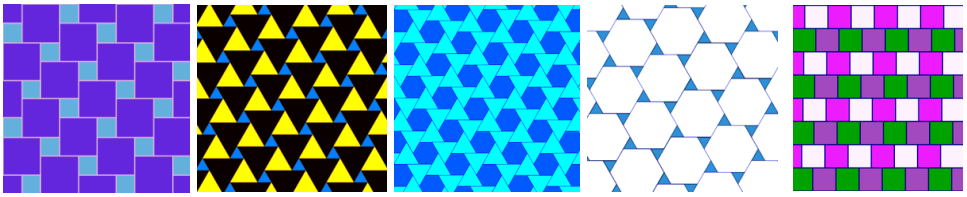


Figure 3.51: Other Ornamental Tilings.

regions, even large ones, that can perfectly be translated to other regions, but the entire plane cannot be copied to itself by translating it in any direction.

Figure 3.52 illustrates the two sets of Penrose tiles, the two rhombi and the kite and dart. The aperiodic behavior results from the special ways in which they are laid, which include matching rules, substitution tiling or finite subdivision rules, cut and project schemes, and coverings. When arranged in any of these configurations, the tiles produce infinitely many different aperiodic tilings. Figure 3.53 illustrates how part of Helsinki's central street is covered by these tiles.

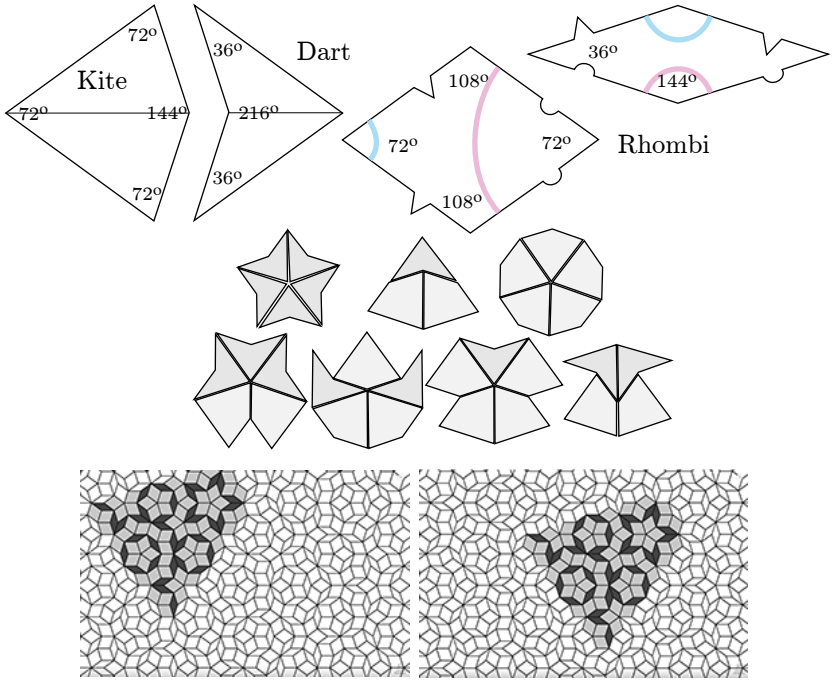


Figure 3.52: Penrose Tilings.

A rigorous proof or this claim is outside the scope of this book, but here is a simple, "hand waiving" explanation that shows how the aperiodicity is due to the use of an



Figure 3.53: Penrose Tilings on a Street.

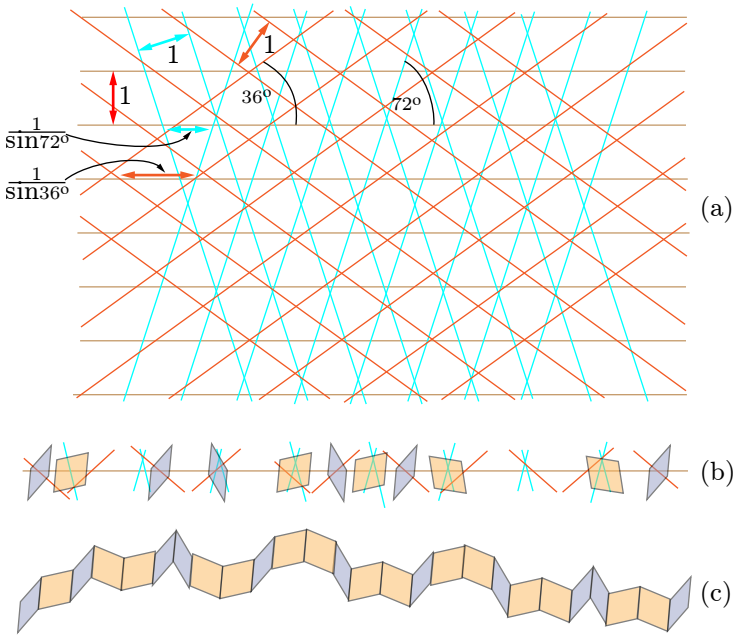


Figure 3.54: Pentagrid and Penrose Tilings.

irrational number, ϕ in this case, but could be any irrational. It starts with a pentagrid, Figure 3.54a, which is constructed as follows:

1. Start with a group of horizontal parallel straight lines, separated by one unit.

2. Create two copies of the group and place them on top of the group after rotating them by 39° and -39°.

3. Create two more copies of the group and place them as before, after rotating them by 72° and -72°.

4. The pentagrid, consisting of five sets of parallel lines, is ready. The figure shows that as we slide along a horizontal line, we cross a $\pm 72^\circ$-line (blue) every $1/\sin 72^\circ \approx 1.05146$ and a $\pm 36^\circ$-line (red) every $1/\sin 36^\circ \approx 1.7013$.

5. Figure 3.54b shows part of an isolated horizontal line with some red and blue lines crossing it. We copy the two Penrose rhombi tiles of Figure 3.52 and place them at the intersections; a thin rhombus at a red line intersection and a wide rhombus at a blue line intersection. The rhombi are placed perpendicular to the horizontal line. The figure illustrates several of those tiles.

6. Finally, Figure 3.54c shows a ribbon of tiles, created by shifting the tiles of Step 4 such that they butt (or juxtapose) one another. A complete Penrose tiling of the plane can be created by performing Steps 4–6 for every horizontal line. Alternatively, parallel ribbons of tiles can be built by placing the first set of lines in an arbitrary direction instead of horizontal.

It is now easy to see why such a finite ribbon, even a very long one, will never match another ribbon anywhere to its left or right. Let's rephrase Step 4 above. As we slide along a a horizontal line, we meet a thin tile every $1/\sin 72^\circ = 1/t$ length units, and we meet a wide tile every $1/\sin 36^\circ = 1/w$ length units. Thus, the densities of the thin and wide tiles in a ribbon are t and w, respectively, and their ratio in the ribbon is t/w or $\sin 72^\circ/\sin 36^\circ$. It is easy to see that this ratio equals

$$\frac{\sin 72^\circ}{\sin 36^\circ} = \frac{\sqrt{\frac{5}{8}+\frac{\sqrt{5}}{8}}}{\sqrt{\frac{5}{8}-\frac{\sqrt{5}}{8}}} = \sqrt{\frac{3+\sqrt{5}}{2}} = \phi,$$

which is the irrational golden ratio. In order for a ribbon to have a match (for its pattern to repeat), the ratio of thin to wide tiles in the ribbon must be a rational number, of the form m/n, with m thin tiles and n wide tiles. Since the ratio is irrational, a copy of a ribbon, even a very long one, will feature slightly different ratios of the two types of tiles and would therefore be different.

This explanation, based on reference [minutephysics 22], is not a proof, but may provide an intuitive understanding of the term aperiodic pattern. Naturally, any irrational number may similarly correspond to the set of aperiodic tiles; there is nothing special about ϕ in this context. Reference [aatishb 23] is a similar explanation, and [spacegoo] is a Penrose tile visualizer. Roger Penrose has patented the tiles in 1975 (U.S. patent 4133152).

The Einstein Tile

Once Penrose published his discovery of the pair of aperiodic tiles, mathematicians started looking for the obvious next step, a set of a single aperiodic tile, a prototile or monotile. The German phrase "ein stein" means one stone, which is why this quest

became known as the Einstein problem, even though it has nothing to do with the famous Albert. The breakthrough came unexpectedly in November 2022, and was finally published in March 2023 [Einstein tile 23]. The developers of this tile named it a hat, and have later shown that it can be generalized to an infinite family of tiles with the same aperiodic property. Reference [Smith 23] has many pictures.

Figure 3.55 Illustrates the simple construction of the hat tile. Start with a regular hexagon and break it into six kite-like pieces. Those are then arranged in an eight-kite configuration to form the hat.

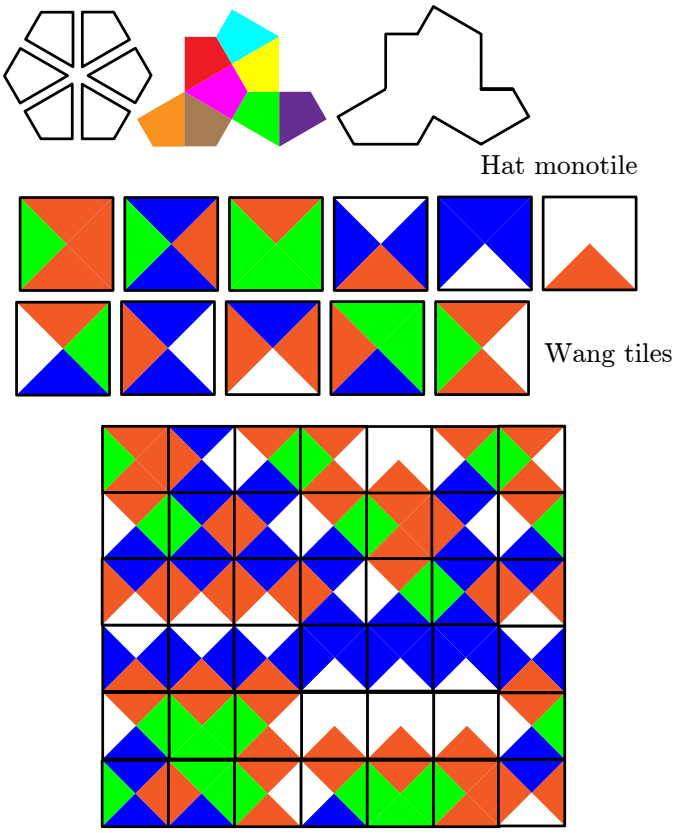


Figure 3.55: The Hat and Some Wang Monotiles.

The Wang Aperiodic Tiles

Wang tiles (also known as Wang dominoes) were proposed by Hao Wang in 1961. The first set of Wang tiles was discovered by Robert Berger in 1964. It consisted of 20,426 distinct tile shapes, a large number that has prompted several researchers to look for smaller sets of Wang aperiodic tiles. The smallest set known today was discovered by Emmanuel Jeandel and Michael Rao in 2015. It consists of 11 tiles and four colors. Figure 3.55 shows the set and a sample of 6×8 tiling. The tiles are laid by matching

the colors of adjacent sides; no rotations or reflections allowed. Laying the tiles is not as easy as it seems, because at each step, the choice of the next tile is very limited. The discoverers have also proved that this set is the smallest possible.

3.8 Tessellations

A tessellation, or tiling, is the process of covering a surface with geometric shapes, called tiles, such that there are no overlaps or gaps between the tiles; they are tightly interlocked. The surface is often two-dimensional (a plane), but can be of higher dimensions. In the simplest case, the tiles all have the same shape, but they can have different shapes. Tessellation is the same as tiling, the subject of Section 3.7, but here we take a practical approach, showing how to create simple tessellations, consisting of repetitions of only one shape of tile.

The first step toward constructing tessellations, either with paper and scissors or with drawing software, is to find out what common geometric shapes tessellate without any modifications. Figure 3.56 shows some obvious choices, followed by less familiar ones. The pattern labeled 6 consists of irregular hexagons and the patterns labeled 5 are irregular pentagons. Notice that circles, as well as polygons with more than six sides, do not tessellate.

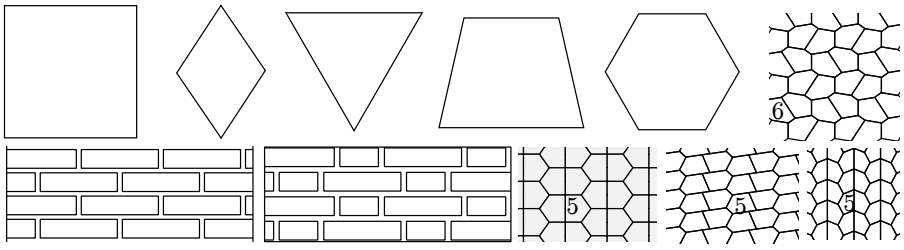


Figure 3.56: Geometric Shapes That Tessellate.

For artistic tessellations, it is common to use either a square, a hexagon, or an equilateral triangle, as illustrated by Figure 3.57. The figure shows four basic techniques to convert one of those shapes to a nontrivial tessellation. The results shown in the figure are not beautiful or interesting, but hard work, persistence, and especially artistic talent, can produce striking effects.

Part 1 of the figure illustrates how easy it is to produce a tessellation by simply translating parts of the original shape. Two curves are drawn along two sides of the hexagon. They are then cut and each is translated to protrude from the opposite side. The resulting shape, which may look very different from the original hexagon, tessellates. With more experience, three adjacent sides of the hexagon can be cut and translated to the opposite side in order to create more complex and interesting shapes.

In part 2 of the figure, circular arcs are drawn inside each face of an equilateral triangle and are then cut, rotated 180° about the red dots, and pasted outside the same

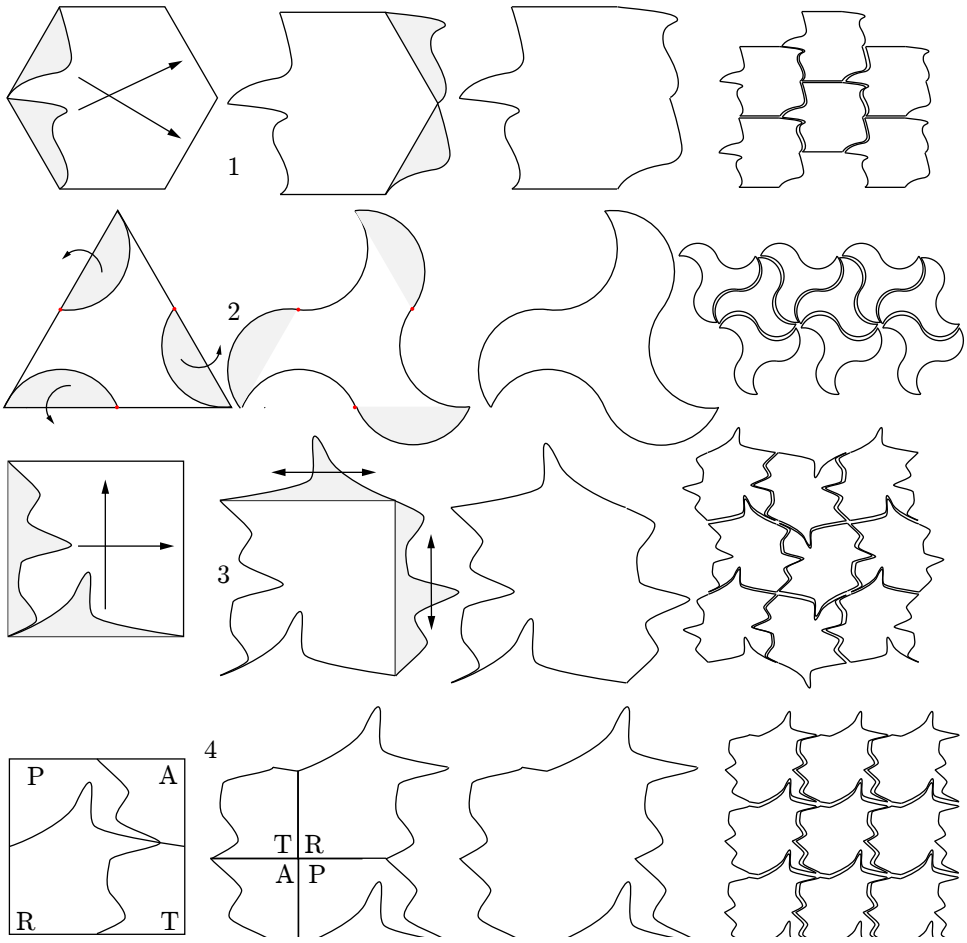


Figure 3.57: Techniques for Artistic Tessellations.

sides. This produces a symmetric shape that looks roughly like a ship's propeller and also tessellates.

Both translation and reflection are applied in part 3 of the figure, which starts with a square. Curves are drawn inside two adjacent sides. Each of the resulting pieces is cut, translated to the opposite edge, reflected, and pasted outside the square. When done accurately, the resulting shape tessellates perfectly, but the individual copies have to be reflected in order to fit together.

The technique of part-trap is demonstrated in part 4. Two curves are drawn in a square. Each runs from the middle of an edge to the middle of the opposite edge. The curves should meet at exactly one point, which partitions the square into four parts labeled `P`, `A`, `R`, and `T`. The parts are cut and are rearranged in the order `TRAP` as shown, to end up with a shape that tessellates.

Any shape obtained by these techniques should be carefully examined by rotating

and reflecting it, looking at it from different directions, and trying to recognize a familiar pattern. If it starts looking familiar, the user can go back, redraw the original curves and go through the steps to recreate a better version. Once satisfied, the result can be further improved by adding texture and colors.

Serious works of art, in the style of Escher or even better, require several steps of cutting pieces and transforming them to other parts of the starting pattern by translation, rotation, and reflection. Figure 3.58 illustrates five steps in the preparation of the famous Escher lizard, and it is obvious that talent is needed, in addition to patience and willingness to experiment.

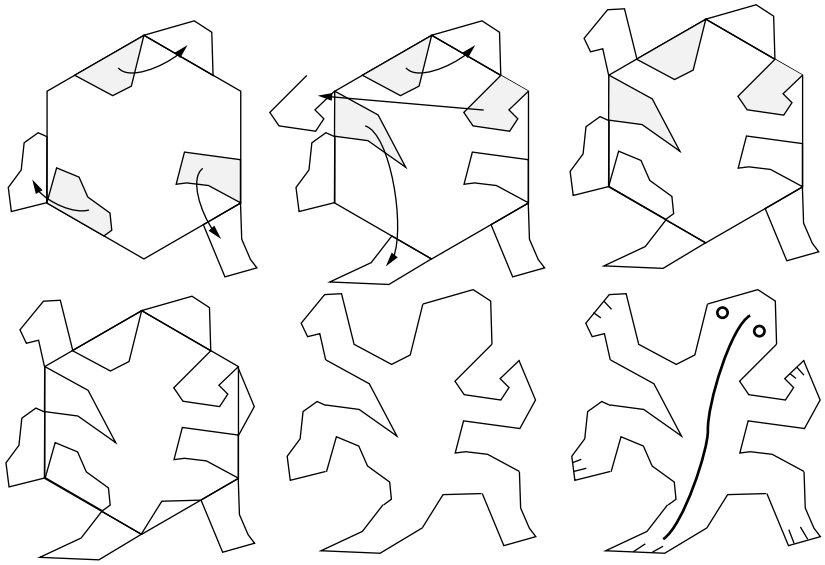


Figure 3.58: Steps in Constructing the Lizard From a Hexagon.

◇ **Exercise 3.25:** Search the Internet for a nontrivial tessellation and summarize its individual steps graphically, similar to Figure 3.58.

3.9 Circle Inversions

The circle inversion discussed here is a two-dimensional nonlinear transformation. The idea is to start with a reference circle in the plane, and invert the entire infinite plane with respect to that circle. Every point P in the circle would map to a point P' outside it, and vice versa. It is easy to decide on the direction of the transformed point P'. The simplest idea is to locate it outside the circle, on the line connecting the center O to P, but what about its distance from O?

Figure 3.59a shows a simple geometric construction where it is clear that triangles OTP and OTP' Are similar, which implies $OT/OP = OP'/OT$ or $OP \times OP' =$

$(OT)^2 = r^2$. Thus, the distance of the transformed point P' from the circle center O is proportional to the inverse of the distance of P from O, and the proportionality constant depends on the radius of the reference circle. Points closer to the center are transformed to points outside the circle and far from it. The center itself is assumed to be transformed to the point at infinity.

Figure 3.59b illustrates an alternative way to compute the transformed point, and part (c) of the figure shows how to compute the transformed point from the original point regardless of whether either is inside or outside the circle. Here are the steps for this simple process:

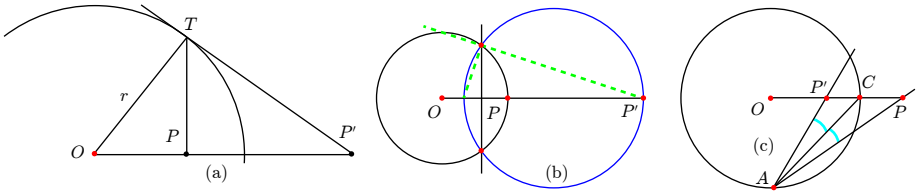


Figure 3.59: Circle Inversion and Similar Triangles.

Select a point P in or out of the circle. Denote the intersection point of ray OP with the circle by C. Connect C with an arbitrary point A, different from C, on the circle. Draw segment AC. Draw ray AP and rotate it such that the two angles (in blue) are equal. This ray now intersects OP at the transformed point P'.

Figure 3.60 illustrates the main properties that make circle inversion useful in practice. The reference circle in the figure is in red.

- In part (a) of the figure, the blue circle passes through the center of the reference circle, which is why its inverse is a straight line (and vice versa).

- The blue circle in part (b) of the figure is not going through the center of inversion, so its image is the cyan circle and vice versa.

- Part (c) of the figure illustrates that if the image of circle A is tangent to circle B, then the image of B (the dashed blue line) is tangent to A.

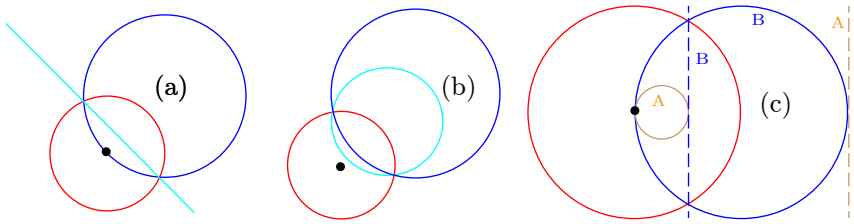


Figure 3.60: Properties of Circle Inversion.

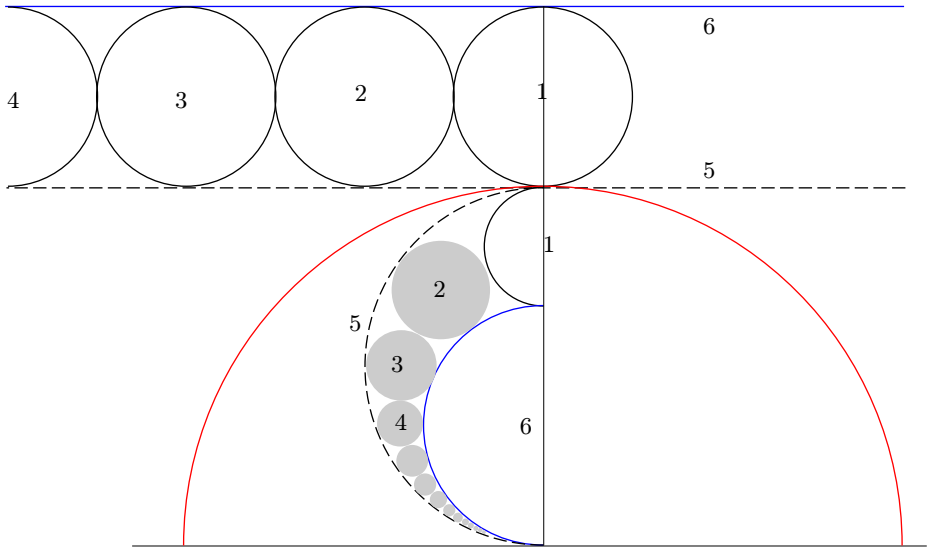


Figure 3.61: Examples of Circle Inversion.

Figure 3.61 illustrates several examples of circle inversion. The reference circle is in red. Notice the three semicircles 1, 5, and 6, whose centers are on the same line. Such a configuration was named an arbelos (Greek for a shoemaker's knife) by Archimedes, who discovered the radius of the unique circle touching all three semicircles.

> Of what use is symmetry?
> —Mao Tse-Tung (Zedong).

3.10 Symmetry in text, speech, and ...

We start with letters. Many letters in the Roman alphabet have mirror symmetries that, being so familiar with the letters, we rarely notice. The letters WAVYTUM|MUTYVAW have left-right symmetries, the letters BDECK have up/down symmetries, and the letters OXIH have both types of symmetry. The letters NSZ have both rotational and left-right symmetries; they are transformed to themselves after a 180° rotation. However, the precise degree of symmetry depends on the font of type.

Symmetric words: The words CHOICE and SWIMS love to be turned both left-right and upside-down, which makes them some of the most symmetrical words in English. The word `dollop` is similar, but slightly less symmetric. MAXIMUM is identical to its mirror image. pod features a 180° rotational symmetry.

Figure 10.5 illustrates how certain digits feature up-down symmetries (but this depends heavily on the font used to print them).

Every letter in the sentence "`HI TIMOTHY, YOU MAY WAX IT NOW`" has vertical symmetry. The same is true for the phrases "MAMA MIA" and "YAMAMOTO."

◇ **Exercise 3.26:** Find English words that begin and end with the letters `und`.

◇ **Exercise 3.27:** The following sentence "I kissed Michael on his ear yesterday." is due to YouTube contributor `Vsauce`. Your job is to find which non-trivial English words (i.e., nothing such as Mmmm or Hmm...) can be placed between any pair of words in the sentence and still keep it grammatically correct.

> Is it odd how asymmetrical
> Is "symmetry"?
> "Symmetry" is asymmetrical.
> How odd it is.
>
> This lovely poem is symmetrical with respect to backward reading.

◇ **Exercise 3.28:** Figure 3.62 Illustrates several numbered parking spaces. Can you guess the number of the occupied space?



Figure 3.62: Numbered Parking Spaces.

A palindrome is a sequence of symbols—such as a word, a number, a phrase, a sound, or a musical piece—that reads or sounds the same backwards as forwards. The term was introduced to English in 1638 by Henry Peacham, who derived it from the Greek $\pi\acute{\alpha}\lambda\iota\nu$ (again) and $\delta\rho\acute{o}\mu o\varsigma$ (way or direction). The basic concept of a palindrome can be traced back to the 3rd-century B.C. We Start with English text examples.

Words: civic, radar, level, rotor, kayak, madam, refer, rotator, deified, racecar, and reviver. The 12-letter onomatopoeia `tattarrattat`, coined by James Joyce in *Ulysses*, is the longest palindrome in English. According to the 2016 *Guinness World Records*, the Finnish 19-letter word saippuakivikauppias (a soapstone vendor), is the world's longest palindromic word in everyday use.

Names: Hannah, Ava, Aviva, Anna, Eve, Bob, and Otto.

Phrases: Dennis and Edna sinned. Lewd did I live, evil I did dwel. No one made killer apparel like Dame Noon. Too bad I hid a boot. Was it a car or a cat I saw? Mr. Owl ate my metal worm. Do geese see God? A man, a plan, a canal, Panama.

A joke. The string ()() is not a palindrome, but ())(is.

Ambigrams: An ambigram is a set of glyphs (letters, numbers, symbols or other shapes) that reads differently when it is inverted, rotated, reflected, or seen from different directions. Most ambigrams are the result of clever artistic calligraphy and are copyrighted. Reference [ambigram 23] has many examples.

Dates: Only the digits of a date are considered in this determination; all other separators (hyphens, slashes, and dots) are ignored. Short digits may be used as in 11/11/11 11:11 or long digits as in 2 February 2020.

A notable palindrome day is 2 February 2020, which is a palindrome in any date convention (yyyy-mm-dd, dd-mm-yyyy, or mm-dd-yyyy) and can be written as either 2020-02-02 or 02-02-2020. This is why this date has been informally declared a "Universal Palindrome Day," a title that also includes 11/11/1111, 12/12/2121, and 03/03/3030.

Speech palindromes. When pronounced in English, the following phrases are palindromic, except perhaps in certain regional accents: "funny enough," "let Bob tell," "new moon," "selfless," "sorry, Ross," "talk, Scott," and "to boot."

◇ **Exercise 3.29:** In your language, find words that rhyme with symmetry.

Palindromes are also found in music, as in the table canon and crab canon. The former is a retrograde and inverse canon meant to be placed on a table in between two musicians, who both read the same line of music in opposite directions. The latter is an arrangement of two musical lines that are complementary and backward.

Numbers: A palindromic number is an integer whose digits remain the same when reversed. It features reflectional symmetry across a vertical axis. Sequence A002113 in the OEIS, reference [Sloane 08], lists some of these numbers.

◇ **Exercise 3.30:** Show that there are infinitely many palindromic numbers in any number base, not just decimal.

◇ **Exercise 3.31:** Show that a palindromic number with an even number of digits is divisible by 11.

Zero is a palindrome by definition.

The first nine terms of the sequence 1, 11, 111, 1111 are palindromes, as also are their squares. Those squares are referred to as palindromic squares, and there are also palindromic cubes (0, 1, and $8 = 2 \times 2 \times 2$ are examples), and palindromic 4th powers. Higher power palindromic numbers are so far unknown.

The 196 Palindrome Quest. Consider the following numeric process. (1) Pick a number, (2) reverse its digits, and (3) add this value to the original number. If the sum is not a palindrome, go back to (2) and repeat. This is known as a delayed palindrome.

Most integers become palindromised by this process in just a few steps, but it is unknown whether this is true for all numbers. The integer 196 is the smallest integer that so far has resisted any attempts to palindromise it, hence the name of this quest.

Antipalindromic numbers. Start with the decimal integer 3802. Convert it to the palindrome 3802083 and subtract each of the added digits 083 from 9 to obtain 3802916, an antipalindromic number. In general, given a palindrome in base b, it is converted to the corresponding antipalindromic number by subtracting each digit of its second half from $b - 1$. This is true for any base. A number can be both a palindrome and an antipalindrome if its base b is odd and each of its digits equals $(b - 1)/2$. The base 5 number 222222_5 is such a number, because $(5 - 1) - 2 = 2$.

Reference [Bergerson 73] brings together over 1,100 different anagrams and nearly as many palindromes, plus a host of related phenomena—charades, circular reversals, vocabulary called poetry, etc.

Continued fractions have been mentioned elsewhere in this book. See, for example, Equation (2.6). It has been proved that if n is any non-square positive integer, the regular continued fraction expansion of $\sqrt{n}$ contains a repeating palindromic block of numbers. The Mathematica command `ContinuedFraction[Sqrt[1205], 37]`, for example, produces the output

```
{34,1,2,2,16,1,12,1,16,2,2,1,68,1,2,2,16,1,12,1,16,2,2,1,68
,1,2,2,16,1,12,1,16,2,2,1,68}
```

where a palindromic string of length 11 appears several times.

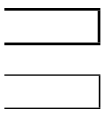
Palindromic number may also occur in mathematical functions and other expressions. A simple example is a polynomial such as $ax^6 + bx^5 + cx^4 + dx^3 + cx^2 + bx + a$, where the coefficients (a, b, c, d, c, b, a) form a palindrome.

Symmetry in graphics. Figure 3.63 illustrates several steps in constructing the well-known space-filling Hilbert curve. There are several ways to explain and construct this curve, and here it is shown as a combination of scaling and rotation. The three-sided cap of the first step is copied, shrunk to half its size, duplicated four times, moved as shown in the figure, and the two bottom copies rotated 90° as shown. The four copies are then connected as illustrated in the figure, and the process can be repeated. The last three steps in the Figure were done with the Mathematica command `HilbertCurve[n]`.

For those unfamiliar with these curves, a space filling curve (several designs are known), is a bijection (a one-to-one mapping) of the real interval $[0, 1]$ to the unit square. Being a bijection, every real number in the interval is mapped by it to some point in the square and every point in the square is the unique map of some point on the line. We intuitively feel that a square has more points than a curve, but the existence of these curves shows that intuition in mathematics, while often being useful, has to be applied carefully.

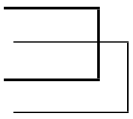
Symmetry in poems. A quatrain is a rhymed poem of four lines or a rhymed stanza of four lines in a longer poem. Many long ballads are written in quatrains, and quatrains also appear in Shakespearean sonnets. Figure 3.64 lists examples of symmetry groupings in quatrains.

I give thee, good fellow, a twelvemonth or twain,
To reach Europe through, from Byzantium to Spain;
But ne'er shall you find, should you search till you tire,
So happy a man as the Barefooted Friar.



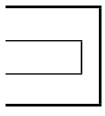
Sir Walter Scott, *Ivanhoe*

The curfew tolls the knell of parting day,
The lowing herd wind slowly o'er the lea,
The plowman homeward plods his weary way,
And leaves the world to darkness and to me.



Thomas Grey

Forgive these wild and wandering cries,
Confusions of a wasted youth;
Forgive them where they fail in truth,
And in thy wisdom make me wise.



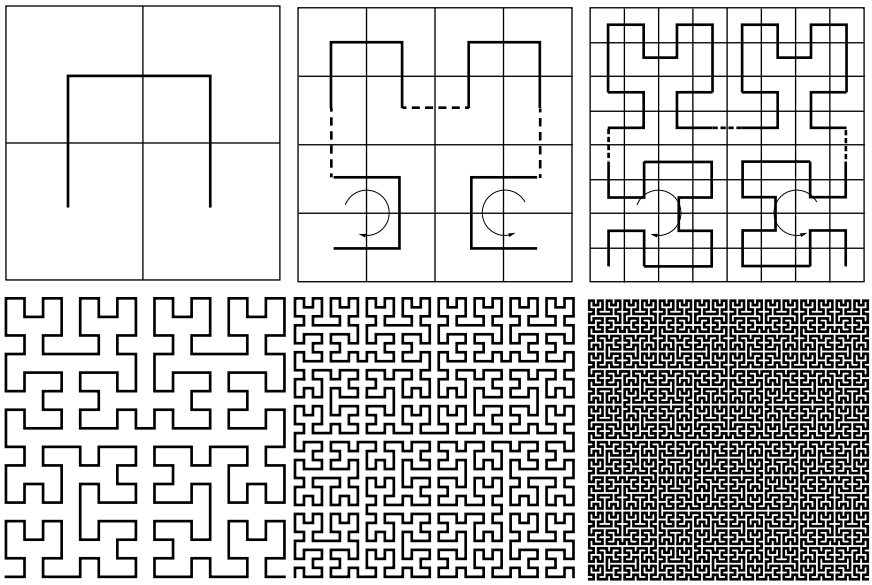


Figure 3.63: Steps in Constructing the Hilbert Curve.

Alfred, Lord Tennyson

Row row row your boat,
Gently down the stream;
Merrily merrily merrily merrily,
Life is but a dream.

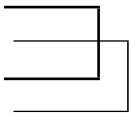


Figure 3.64: Symmetry in Rhymes.

Symmetry: A Poem by `chatGPT 4o`, May 2024.

In nature's art and crafted forms,
Symmetry appears, and beauty swarms,
A balance struck in line and arc,
In every snowflake, tree, and park.

Reflections in a mirrored lake,
A perfect stillness they create,
In butterfly wings, a balanced sight,
Symmetry, a pure delight.

In crystals' lattice, tight and neat,
Symmetry repeats, so sweet,
Each angle, edge, and face aligned,
A dance of atoms, finely timed.

In algebra, a concept blooms,
Symmetry groups in ordered rooms,
With operations, clear and bright,
They map out shapes in purest light.

From cyclic twists to dihedral turns,
In every structure, order yearns,
A square rotates, reflections blend,
Symmetry groups, rules transcend.

The group of Klein, a fourfold grace,
With flips and spins in mirrored space,
Each element, a transformation,
A testament to calculation.

In higher realms where spaces bend,
Symmetry's rules still extend,
In n-dimensions, forms evolve,
With symmetry, problems resolve.

The universe, in grand display,
Shows symmetry in vast array,
From galaxy spirals, vast and wide,
To particle spins, in quantum tide.

Noether's theorem, deep and grand,
In physics, takes a bold stand,
Each symmetry, a conserved law,
In nature's code, a flawless draw.

In art, in music, patterns found,
Symmetry's echoes all around,
From Da Vinci's lines to Bach's sweet tunes,
Symmetry sings in varied runes.

In Escher's worlds, where fish take flight,
Symmetry twists in paradox sight,
With tessellations, infinite,
A play of form, exquisite, bright.

In every face, a trace we see,
Of symmetry's soft decree,
From human forms to shells at sea,
A universal harmony.

So let us marvel at this theme,
In symmetry, a dreamer's dream,
For in its balance, we discern,
A deeper truth at every turn.

Symmetry groups, with elegance,
Show math's purest resonance,
A language of the universe,
In every line, its verses terse.

In fractals' endless self-embrace,
Symmetry finds another place,
Recursive, infinite, it stands,
A testament to nature's plans.

Thus in the world, both near and far,
Symmetry shines, a guiding star,
A principle, both grand and small,
The hidden code, the essence of all.

Symmetry looks good to us; we want more of it.
—Susan Messing.



4
Infinity

Frequently we encounter or come across statements like the following: "after spending hours walking through the seemingly endless hallways of the (Louvre, Pentagon, Paris sewer system), we eventually reached our destination with sore feet." In our daily lives, we are aware that those hallways, no matter their location, were not endless, but rather extremely lengthy, with "infinite" being an exaggerated term for emphasis on their great length.

The concept of infinity, whether it exists or not, its significance, and the consequences it brings, has been a subject of wonder, contemplation, and dialogue for centuries. When early humans started observing their surroundings and pondering how things functioned, they began exploring the concepts of infinity in terms of both time and space. Has the world existed forever? Will it continue to exist indefinitely? Does it have unlimited size? Imagine if we continued in a straight line. Is it possible to ever reach the end of the universe, or is the journey to get there never-ending? What if we begin with an object, divide it in half, then divide one of those halves in half again and continue this pattern? Can this process theoretically be continued indefinitely? Therefore, infinity pertains to both space and time, as well as to both the extremely large and the extremely small. Is it possible for a particle of matter to be infinitesimally tiny? Is it possible for there to be infinitesimally smaller units of time?

Aristotle came up with an interesting way to approach infinity. He argued that humans will never be able to grasp infinity as a concept or as an object and can only hope to approach it in steps. Thus, we cannot conceive of the complete, infinite set of integers, but we can approach this set by realizing that given any finite set of integers we can easily extend it to a larger, finite set. (This argument makes sense when we realize that during the whole of human history, in all the calculations ever made, only a finite number of integers have ever been generated and examined. Even in the future, with the help of ever faster computers, only a finite set of integers will ever be generated.) Thus, infinity is only potential. When dealing with an infinite set, wherever we are in

the set, we can always go further. We will never reach the end. Infinity is not a quantity that can be completed.

Nothing that we can see, measure, or perceive in the universe around us is infinite. Even the farthest galaxies are a finite (albeit very large) distance from us. Thus, it seems that the only infinite quantities we deal with are philosophical and mathematical concepts and objects, and perhaps also the universe itself. It is not currently known whether the universe is infinite, which is why any experiences we have with infinity come from mathematics and from thought experiments. Many mathematical sets, quantities, and objects are infinite. There are infinitely many integers, infinitely many odd integers, many infinite sequences of numbers (in fact, infinitely many infinite sequences), many infinite series that converge to finite values, and many expressions, functions, and sets that generate infinite patterns.

> Two things are infinite, as far as we know, the universe and human stupidity.
> —Attributed to Albert Einstein by Frederick S. Perls.

4.1 A Short History of Infinity

Perhaps the earliest writings about infinity came down to us from the ancient Greek philosophers, chief among them are:

Anaximander, who claimed that infinity is *the* founding principle of reality. From it are created an infinite number of worlds that fill the entire universe.

Heraclitus, in contrast, believed that time is infinite and it is through its infinity that we perceive our own existence.

Zeno of Elea came up with several famous arguments demonstrating the paradoxical nature of the infinite by closely examining and analyzing everyday actions such as running or shooting an arrow.

But it was Georg Cantor, in late 19th century, who applied modern mathematical methods and techniques of proofs to the study of infinite sets, and came up with the revolutionary notion of different infinities. His creation, beginning in 1874, of set theory, has since been an endless source of debate, arguments, and research.

Finally, another approach to infinity, that of recursive mathematical processes that can also be demonstrated graphically, has been advanced and promoted by Benoit Mandelbrot, who coined the term fractal (Section 8.1) and has brought to the attention of the mathematical world the names of pioneers in this field, names such as Sierpinski (carpet and gasket fractals), Peano (space-filling curve), Menger (sponge), Julia (set), and Harter and Heighway (dragon curve).

> The infinite we shall do right away. The finite may take a little longer.
> —Stanislaw Ulam.

4.2 Mathematical Infinity

Even though we cannot grasp infinity, it is often easy to deal with it in mathematics. First, we present simple arguments to convince the reader that infinity is not a number. Given any number a, we can always add 1 to it, ending up with a number $b = a + 1$ that is greater than a. Now imagine adding 1 to infinity. The quantity $(\infty + 1)$ cannot be less than infinity. It can only be greater than or equal to infinity. If we assume that $(\infty + 1)$ is greater than infinity, then we have a new, greater infinity, to which another 1 can be added, to form an even greater infinity, and so on. This contradicts the concept of infinity which is "beyond anything." We therefore have to accept that $(\infty + 1)$ equals infinity, which implies that infinity is not a number. Similarly, if infinity were an integer, then it must be odd or even. If infinity is odd, then $2 \times \infty$ would have to be even, but we cannot accept that $2 \times \infty$ is another infinity, greater than the "original" infinity. Therefore, $2 \times \infty$ must equal infinity and so must be odd, a contradiction.

Thus, infinity is not a number, not even a very large number. It is not simply a little bigger than any number we can think of. It is a different concept and should be considered differently and treated differently from numbers.

We follow with several examples that illustrate the usefulness of infinity in mathematics.

Example: The concept of a *limit* is central to calculus. The limit of the function $f(x) = 2x$ when x approaches infinity is infinite. This is denoted by $\lim_{x\to\infty} 2x = \infty$. Similarly, $\lim_{x\to\infty}(x + 1)$ is infinite, but the limit of the ratio $2x/(x + 1)$ is the finite value 2.

Example: The set P of all positive integers is infinite, and so is the set E of all the positive even integers. It is intuitively clear that E cannot be greater than P, because E is a subset of P. Thus, E can only be smaller than or equal to P, and the surprising fact is that we can show that these two infinite sets are equivalent in the sense that their elements can be matched. An element p of P is a positive integer and is matched with $2p$, which is even and therefore an element of E. Conversely, an element e of E is an even positive integer, so $e/2$ is a positive integer and therefore an element of P.

Set theory, the creation of Georg Cantor in the 1870s, shows that the set of the rational numbers is equivalent to the set of all the integers. This is surprising because there are so many "more" rational numbers than integers. This result shows that in principle, an integer tag (or subscript or index) can be assigned to every rational. However, set theory also shows that the set of all real numbers is infinitely greater than the set of integers, even though both are infinite. Thus, there is no way to assign an integer subscript to every real number, even though there are infinitely many integers. An unintuitive result.

Example: Computing the value of the infinite continued fraction

$$1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \ddots}}}}$$

seems hopeless, but turns out to be simple. We denote this sum by ϕ and examine the expression $1 + 1/\varphi$. A little thinking should convince the reader that this sum is

identical to the original φ because this continued fraction is infinite. Thus, we end up with the unusual quadratic equation $1 + 1/\varphi = \varphi$ whose solutions are $(1 \pm \sqrt{5})/2 \approx (-0.61803398875, 1.61803398875)$. Equation (2.7) and Subsection 2.11.4 make it clear that this φ is the golden ratio.

This quadratic equation is unusual because it can be written in the forms $1/\varphi = \varphi - 1$ and $\varphi + 1 = \varphi^2$. The former form implies that inverting the unknown number φ produces the same result as subtracting 1 from it. The latter form tells us that adding 1 to φ yields φ^2.

Example: The integers are discrete. Each integer has a successor and a predecessor, but the real numbers are dense everywhere. There are infinitely many real numbers. There are also infinitely many real numbers in the short interval $[0, 1]$. In fact, there are infinitely many real numbers in any interval, even the shortest one that we can imagine. Thus, even though there are infinitely many integers and infinitely many real numbers, we may not find it so difficult to accept that the reals constitute a different magnitude of infinity and that there even are many different infinities (perhaps even infinitely many).

Example: Consider the limit of the infinite sequence of nested square roots

$$\lim_{x\to 0^+} \sqrt{x + \sqrt{x + \sqrt{x + \cdots}}}. \tag{4.1}$$

We show why this limit equals 1. As x shrinks and approaches zero, this expression becomes the limit of square roots of zeros. As long as we consider only a finite number of nested square roots, its limit in x indeed converges to zero. To make this clearer, we can write Equation (4.1) in the form

$$f(x, n) = \underbrace{\sqrt{x + \sqrt{x + \sqrt{x + \cdots}}}}_{n}. \tag{4.2}$$

For any finite value of n, the limit of Equation (4.2) is a limit of zeros, so it is also zero.

However, when we extend the limit of n to infinity, Equation (4.1) becomes a double limit whose inner limit (of x approaching zero), is surrounded by an outer limit of infinite n square roots. Thus,

$$\lim_{n\to\infty} \lim_{x\to 0^+} \sqrt{x + \sqrt{x + \sqrt{x + \cdots}}}. \tag{4.3}$$

When we now reverse the order of limits of Equation (4.3), its sum changes. The new double limit

$$P = \lim_{x\to 0^+} \lim_{n\to\infty} \sqrt{x + \sqrt{x + \sqrt{x + \cdots}}}, \tag{4.4}$$

has an inner part that consists of an infinite number of square roots, so in order to compute the outer limit (on x), we label the sum of expression (4.4) by P, and use our previous technique to write $P = \sqrt{x + P}$, which yields the quadratic equation

$$P = \frac{1 \pm \sqrt{1 + 4x}}{2}.$$

Since x goes down to zero, it must be positive, implying that P must also be positive, so we pick the positive solution of the quadratic equation, which is $P = 1$.

> Most human beings have an almost infinite capacity for taking things for granted.
> —Aldous Huxley.

4.3 Potential and Completed Infinities

When dealing with, or thinking of, infinity, we normally treat it as a potential quantity. Treating infinity as a completed quantity leads to paradoxes, so we avoid it in all cases except in set theory, where a completed infinity produces new, unexpected, and unusual results that many find beautiful and deep. The following examples show what paradoxes may arise when we assume (often implicitly) that infinity can be completed.

Example: Given two numbers a and b, their average is $(a + b)/2$. Imagine an infinitely long, vertical line on which the positive integers are written, starting with 1 at the bottom and going all the way to infinity at the top of this line. What would be a typical average positive integer? This is like asking what number appears in the middle of this line. The answer is that there is no such number. As we climb this infinite vertical line, we always remain at the same distance from the top, because the top does not exist; the line has no end. No integer is a typical average between zero and infinity. To see why this is true, pick the largest integer you can think of and denote it by a. This number is not an average integer because the distance between it and the bottom of the line is the large but finite quantity a, while the distance between it and the top of the line is infinite.

This inability to specify an average integer is not a result of our ignorance or lack of scientific ability. An average integer cannot be found because such an integer simply does not exist. This sounds like a paradox, because there are so many integers, but the paradox is resolved when we realize that the words "going all the way to infinity at the top of this line" imply that there is a top. In stating this example, we implicitly assumed that it is possible to reach the end of an infinite line. We assumed that infinity is a completed quantity.

Example: This is referred to as Albert of Saxony's paradox. In many places around the world, the construction of a house starts with a frame consisting of wooden beams; lumber. Builders often use phrases such as "a 2×4" or "a long 4×4." Now imagine a long 1×1 beam. It is 1 inch high and 1 inch wide and as a thought experiment we extend it to infinite length. The beam starts at our feet and continues until it disappears from sight. With a powerful telescope we can see it until it disappears beyond the orbit of the moon. Even though this beam is infinitely long, it does not occupy an infinite amount of space. It is quite small and narrow, we can step over it, move it, and cut it.

Now imagine this beam cut into infinitely many small cubes, with dimensions of $1 \times 1 \times 1$ each. We pick the first cube and arrange the next 26 cubes around it to form a larger, $3 \times 3 \times 3$ cube consisting of $3^3 = 27$ small cubes. Next, we pick the next $5^3 - 3^3 = 625 - 27 = 98$ small cubes and arrange them around the large cube to increase its dimensions to $7 \times 7 \times 7$. There is an infinite supply of small cubes, so this

process increases the large cube more and more without limit. The original thin, long, and unobtrusive beam can in this way supply building material for a cube that fills up the entire universe (or at least the observable universe). This is again a paradox, arising from our implicit assumption that the *entire* infinitely-long beam can be cut into small cubes.

Example: The infinite grand hotel (this entertaining concept seems to have originated with the eminent mathematician David Hilbert, see [Grand-Hotel 11], but was popularized by [Gamow 61], pp. 17–18.). At the time of writing (late 2011), the largest hotel in the world (by number of units) is the Izmailovo hotel complex, located in the Izmaylovo District of Moscow. (It was built for the participants and visitors of the 1980 Moscow Olympics.) It features four buildings with a total of 7500 rooms. A guest arriving at this hotel without a reservation, even at the height of season, is unlikely to be turned away, but in principle, even this hotel may be full, in which case, the next guest will have to find a room elsewhere. The second largest hotel in the world is currently the Venetian and The Palazzo complex in Las Vegas, Nevada, boasting 7,117 rooms.

Vast as they are, these hotels are dwarfed by the grand hotel envisioned by Hilbert, because this hotel is infinite. It contains a countable infinity of rooms (the term countable implies that the rooms can be numbered, each has an integer number). It is easy to show that this hotel has a number of surprising, counterintuitive properties that make it especially attractive. The most important such properties are listed here:

- Imagine the case where the infinity grand hotel is full; every room is occupied (contrary to what you might think, this is common). A new guest (without a reservation) arrives late at night and is asking for a room. What is the manager to do? In any other hotel, the guest would simply be turned away and sent to another big hotel (Izmailovo?), but in the infinity hotel one more guest can always be accommodated. All that the manager has to do is move the current occupant of room 1 to room 2, the occupant of room 2 to room 3, and so on. Once this is done, room 1 is empty, and can be prepared for the new guest. Easy! This encouraging result is equivalent to saying that $\infty+1=\infty$.

- It is really late. The tired manager prepares to go to sleep when all of s sudden a bus arrives with an infinite number of guests. Clearly, no other hotel can accommodate so many guests, especially when it is already full, but our infinity grand hotel is different. The manager can find rooms for any number of new guests, even an infinite number (as long as it is countable). Here is how: Simply move the occupant of room 1 to room 2, that of room 2 to room 4, and in general, from room n to room $2n$. All the rooms with numbers of the form $2n$ (even) will now be occupied and all the rooms with numbers $2n+1$ (odd) will be empty. All the (infinitely many) new guests can now be accommodated.

- In the morning, the tired manager is woken up by irate guests. "As if it is not enough that I was moved once because of the unexpected new guest, I was moved again (and to an even-numbered room) when this huge bus arrived. This is unacceptable; I quit. Tonight I will try the Izmailovo hotel. They will surely will let me spend a quiet night." The result is that half the guests, those in the even-numbered rooms, check out, leaving the hotel with only 50% occupancy. The poor manager is worried about his job, but then an angel appears, in the form of his little girl (I always said she was an angel). "Don't worry papa, we still have an infinite number of guests in the odd-numbered

rooms. Let's compact them and we'll go back to 100% occupancy." She also explains how to do this compaction. Move the occupant of room 3 to room 2, the occupant of room 5 to room 3, room 7 to room 4, and in general, room $2n+1$ to room $(n+1)$. The result is a full hotel and a manager who keeps his job. As an added bonus, the occupant of room 1 does not have to be disturbed.

■ The ultimate in performance occurs one day, when infinitely many tourist buses arrive, each with infinitely many weary travelers. They all want rooms even though the hotel is, as you guessed, full. The first part of the solution is to move current occupants, as before, so as to have all the odd-numbered rooms vacant. The second part is to assign each new arrival a (positive integer) number. We cannot simply number the occupants of bus 1, followed by those of bus 2, and so on, because each bus has infinitely many passengers. Perhaps the simplest method is the infinite square, originally used by Cantor to prove the countability of the rationals. (1) Assign each new arrival a pair of numbers (bus number, passenger number). (2) Arrange the pairs in an infinite square matrix where each row corresponds to a bus and each column corresponds to a passenger. (3) Scan the matrix in zigzag as illustrated in Figure 4.1. This assigns a unique number (positive integer) to each passenger, and now passenger n can check in and is accommodated in the odd-numbered room $2n+1$.

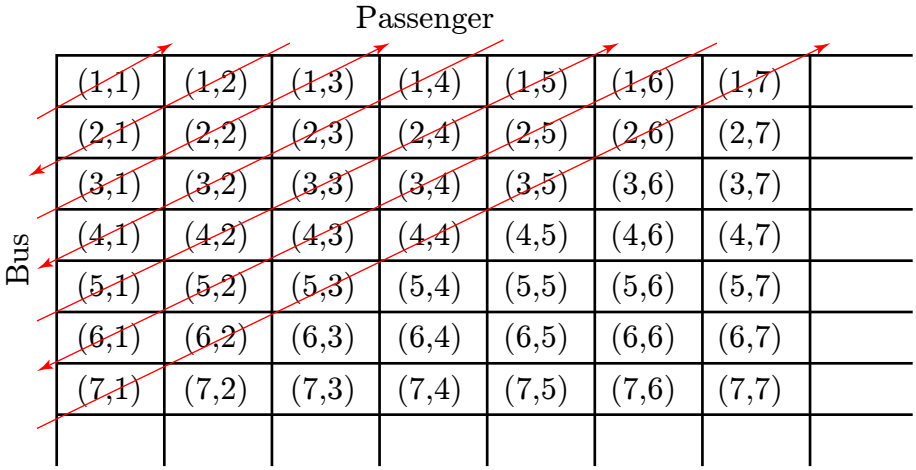


Figure 4.1: Scanning in ZigZag.

■ In addition to accommodating any number of guests, the grand infinity hotel has many other advantages (and features paradoxes), not found in conventional, finite hotels. Here is one example. A big sign in the lobby says No Smoking. Even more, no cigars may be brought into the hotel. Clever guests can circumvent this rule and smoke even though none has to bring cigars. The (infinite) smoking process starts when the occupant of room 1 goes to room 2 to borrow a cigar. The occupant of room 2 goes to room 3 and borrows TWO cigars. He smokes one and gives the other one to the guest in room 1. In general, the person in room n borrows n cigars from room $(n+1)$, smoke one cigar,

and give the remaining ones to the occupant of room $(n-1)$. Thus, all the guests break the no smoking rule, even though none has brought any cigars into the hotel. Where did the cigars come from? This is a paradox similar to the ones discussed earlier.

Many find this type of hotel very entertaining, and reference [Grand-Hotel 11] lists several works of fiction where it appears.

> We cannot speak of infinite quantities as being the one greater or less than or equal to another.
>
> —Galileo.

Infinite series and sequences are discussed elsewhere, but here we look at one baffling infinite series, known as Grandi's series, that is divergent. It has no well-defined sum, although its partial sums do not increase without a limit. It is defined as the infinite sum

$$S=\sum_{n=0}^{\infty}(-1)^n=1-1+1-1+1-1+1-1+1\cdots.$$

In contrast with many other infinite series, our S is easy to evaluate, but it is also easy to show that it can have any value whatsoever. First, we show that it equals 0, simply by placing parentheses in a special way

$$S=(1-1)+(1-1)+(1-1)+(1-1)+\cdots=0.$$

Changing the positions of the parentheses indicates that $S=1$

$$S=1+(-1+1)+(-1+1)+(-1+1)+(-1+1)\cdots=1.$$

Another change produces the unexpected sum $S=1/2$

$$S=1-(1-1+1-1+1-1+1\cdots)=1-S\rightarrow S=1/2,$$

(see also note on Page 6.) This strange behavior is another result of treating infinity as completed.

> I am so in favor of the actual infinite that instead of admitting that Nature abhors it, as is commonly said, I hold that Nature makes frequent use of it everywhere, in order to show more effectively the perfections of its Author.
>
> —Gottfried Wilhelm Leibniz in a letter to Simon Foucher, 1692.

4.4 Unexpected Results of Infinity

- **Ant on a flexible string**

This example of a beautiful, unintuitive result that relies on the harmonic series is described here is following the video at [yt.ant.string 23]

Imagine a string of length L with infinite stretchability. It can be stretched without limit and without breaking. We place an ant at a point l on the string. The ant moves forward continuously at a low speed of, say, 1 cm/hr, while the string is stretched at a much higher speed, say, 1 km/hr. Intuitively, it feels as if the ant is going to get behind more and more all the time, but it is easy to show that the ant would actually move forward, albeit slowly, and will eventually reach the end of the string no matter the ratio between its speed and the stretching speed.

To see why this is so, we should think of the problem in terms of the distance remaining for the ant to cover. Consider the simple case where the ant starts at the middle of the string (Figure 4.2), and is initially at rest. When the string is stretched, it is stretched equal amounts in front of the ant and behind it, so even if the ant is at rest, it moves with the middle of the string. It now seems possible that if it is also moving by itself, at any speed, it would continuously gain distance and may eventually reach the end of the string, however long it takes.



Figure 4.2: Ant on a Stretched String.

This still has to be proved rigorously, but the analysis is simple. Let's run the process second by second and assume that the ant starts at the beginning of the string. We denote the initial length of the string, the distance it is stretched each second, and the distance the ant moves each second by L, D, and d, respectively. After the first second, the ant has covered distance d, and the string is $L + D$ units long.

During the kth second, the ant covers d units of the string whose length (at the end of the second) is $L + kD$. Thus, the part of the string covered by the ant in the kth second is the ratio $d/(L + kD)$ and the sum of the parts the ant covers in the first n seconds is

$$\sum_{k=1}^{n} \frac{d}{L + kD}. \tag{4.5}$$

However, it is easy to see the following relations

$$\frac{d}{L + kD} \geq \frac{d}{kL + kD} = \frac{d}{L + D}\left(\frac{1}{k}\right),$$

which is why

$$\sum_{k=1}^{n} \frac{d}{L + kD} \geq \sum_{k=1}^{n} \frac{d}{L + D}\left(\frac{1}{k}\right) = \frac{d}{L + D}\sum_{k=1}^{n}\left(\frac{1}{k}\right).$$

However, the latter sum of $1/k$ is the well-known harmonic series, which has long been proved to diverge, which proves that the sum of Equation (4.5) also diverges, so in the limit, where n approaches infinity, the ant has covered all the parts of the stretched string and has reached its end, no matter how long it is.

- **Algebraic? transcendental?**

  The well-known Euler constant can be represented by the infinite sum

$$e = \sum_{k=0}^{\infty} \frac{1}{k!},$$

(See also Page 130) all of whose terms are algebraic, but because of the mysterious, unpredictable nature of the infinite, it was possible for Charles Hermite to prove, in 1873, that no algebraic equation yields e as a solution. It is transcendental even though each of its components is algebraic.

It's easy to find more examples, as the following illustrates.

The following two-player game is a representative of an entire class of thought processes that illustrate the subtle nature of the infinite. The game is played in steps where in each step, each of the two players, A and B, receives the next two integers, appends them to his collection, and deletes one number from that collection. Player A deletes the smallest number he has, and player B deletes the largest number in his collection.

In step 1, each player receives the integers 1 and 2. Player A deletes 2, while B deletes 1. In step 2, each receives 3 and 4. Player A now has 1, 3, and 4, so he deletes the 4, while player B, who has 2, 3, and 4, deletes the 2. In step 3, each receives 5 and 6. Player A deletes the 6, retaining 1, 3, and 5, while B deletes the 3, keeping 4, 5, and 6. The table lists the first three steps.

A: 1 ~~2~~ 3 ~~4~~ 5 ~~6~~ ...
B: ~~1~~ 2 ~~3~~ 4 5 6 ...

As the game progresses, A retains more and more odd positive integers (n such integers after step n), while B keeps longer and longer sequences of the form $(4, 5, 6)$, $(5, 6, 7, 8)$, $(6, 7, 8, 9)$, and so on (also n integers after step n).

The question is, what will the two players have at the end of this game? The case of player A seems straight. He will simply have *all* the positive odd integers. The case of B, however, is not that clear. On one hand, B has n integers after step n, so at the end he will simply have infinitely many integers. On the other hand, player B deletes the integer 1 in step 1, the integer 2 in step 2, and so on, so at the end of this game he will have deleted *all* the positive integers that he received and should be left with nothing! Again a paradox, and again it is resolved when we realize that the words "at the end of this game" at the beginning of this paragraph are meaningless. This game is infinite, so it will not have an end.

In a variant of this game, each player is given all the positive integers when the game starts. Player A deletes all the even integers and is left with the odd ones. This is a case where infinitely many integers input and infinitely many integers output leaves us with infinitely many integers. Player B, on the other hand, deletes the integer n in

step n, so is eventually left with nothing. In this case, infinitely many integers input and infinitely many integers output leaves the player with nothing (no integers).

This is a paradox, and by now it should be obvious that the resolution is Beware of the infinite. The game would never end, which is why it is meaningless to consider its end result.

Here is a variation known as the Ross-Littlewood paradox. A jar contains a (countable) infinity of balls. In each step we remove 10 balls and put back one of them. After n steps, a total of $9n$ balls would have been removed, but what would be left after infinitely many steps? On one hand, the jar should be empty, because removing $10-1=9$ balls at a step would remove nine infinities of balls after infinitely many steps. On the other hand, we put back one ball in the jar at each step, so after infinitely many steps it should have infinitely many balls; the same number as at the start.

To make this process more dramatic, we can speed up the time. Assume that we remove the first 10 balls at 11 pm, the next 10 balls 30 minutes later. The next batch after another 15 minutes, and so on, continuing faster and faster until midnight. (This implies infinitely many steps, without having to mention the word infinity.) We then go to bed and wait until morning to take a peek at the jar. Surprise!

Another variation is the Thomson's lamp, an old paradox/puzzle related to Grandi's series of Page 246. Given a lamp that is controlled by a flip switch, start turning it on and off faster and faster. The question is what would its state be after an infinite number of switches? The resolution of this paradox is the same as for its predecessors, namely there will not be an end. The state of the lamp is well defined after any finite number of switches, but not after an infinite number, even if this infinite process accelerates and takes only a finite period of time.

It is also possible to explain this paradox in two other ways. (1) Current physical theory cannot describe timescales shorter than the Planck time (approximately 5.39×10^{-44} sec.) Thus, physics makes it impossible to perform the experiment and it suggests that its final outcome is nonexistent because of the way our world is constructed. (2) Mathematics claims that the last step does not exist since the sequence of steps is infinite.

A variation of this paradox states that the Thomson switching process starts a minute before midnight and lasts for one minute. The switching occur at 11:59, 11:59:30, 11:59:45, and so on. What is the state of the lamp at midnight? Here we claim that every step in the process takes place before midnight, so there is no step at precisely midnight. See also the racecourse paradox on Page 319 and Achilles and the fly paradox on Page 325.

- Sierpinski-Mazurkiewicz Paradox

Imagine a set S of points in the real plane. Each point in S is a pair (x,y) of real numbers. Partition S into two disjoint subsets A and B whose union $A\cup B$ is the entire S. Now try to come up with a transformation that transforms A into the entire set S. We start with A, which is a subset of S, apply the transformation to every element of A, and end up with S, a bigger set! Let's make it even harder. Come up with another transformation that transforms B into the entire set S. This sounds impossible, but because of the nature of the infinite, such a set and such transformations exist, and the example described here, based on reference [Su 25], is known as the Sierpinski-Mazurkiewicz paradox.

To appreciate this unusual set of transformations, it is important to understand the relation between points on a plane and polynomials. One of the many well-known equations attributed to Euler is

$$e^{i\theta} = \cos(\theta) + i\sin(\theta), \tag{4.6}$$

from which we derive the special case $e^i = \cos(1) + i\sin(1) \approx 0.54 + 0.84i$. This complex number can be visualized as a point $p = e^i$ on the complex plane, from which we conclude that any polynomials in p, such as $p + 1$ and $3p^2 - 2p + 4$, are also points. In general, any polynomial of the form $\sum a_i p^i$ with real coefficients a_i and with $p = e^i$ is associated with a unique two-dimensional point.

Another piece of knowledge that is needed here is the rotation of complex numbers. Given the general complex number $(a + bi)$, we rotate it by multiplying it by the counterclockwise rotation matrix to obtain

$$\begin{aligned}
&\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a\cos(\theta) - b\sin(\theta) \\ a\sin(\theta) + b\cos(\theta) \end{bmatrix} \\
&= \big(a\cos(\theta) - b\sin(\theta)\big) + i\big(a\sin(\theta) + b\cos(\theta)\big) \\
&= (\cos(\theta) + i\sin(\theta))(a + bi) \\
&= e^{i\theta}(a + bi).
\end{aligned}$$

Thus, a counterclockwise rotation of the complex number z is the product $e^{i\theta}z$, and a clockwise rotation is similarly $e^{-i\theta}z$.

Armed with this knowledge, we choose as our S the set of all polynomials in $x = e^i$ with nonnegative integer coefficients. This set is unbounded, because the integers are unbounded, but it is countable, because the coefficient are integers.

Subset A will be those polynomials with a constant term of zero, and subset B will be the polynomials whose constant term is positive. Thus, an element of A may be $4x^3 + 6x^5 = x(4x^2 + 6x^4)$ (a multiple of x), while an element of B may be $4x^3 + 6x^5 + 7$. Remembering that $x = e^i$, we apply the following transformations:

- Translate the elements of B by subtracting 1 from each. The example above becomes $4x^3 + 6x^5 + 6$. This produces the entire set S, because those elements of the new B whose constant term is positive remain elements of B, while those elements of the new B whose constant term is zero become the elements of the new A.

- Rotate all the elements of A clockwise by one radian. Based on the rotation discussion above, rotating the point $x = e^i$ clockwise by one radian is done by multiplying by $e^i \times e^{-i} = 1$. Since the elements of A are multiples of x, such a rotation converts an element of A (with a constant term of zero) to a general element of S (with a constant term of zero or positive).

◇ **Exercise 4.1:** Figure 4.3a shows a circle of diameter 1 surrounded by a square of side 1. The circumference of the circle is π and the circumference of the square is 4. Parts (b) and (c) of the figure illustrate how the circumference of the square can be fragmented in steps while maintaining its total length. As long as the square is fragmented a finite number of times, it circumference stays the same 4, but in the limit, it becomes identical

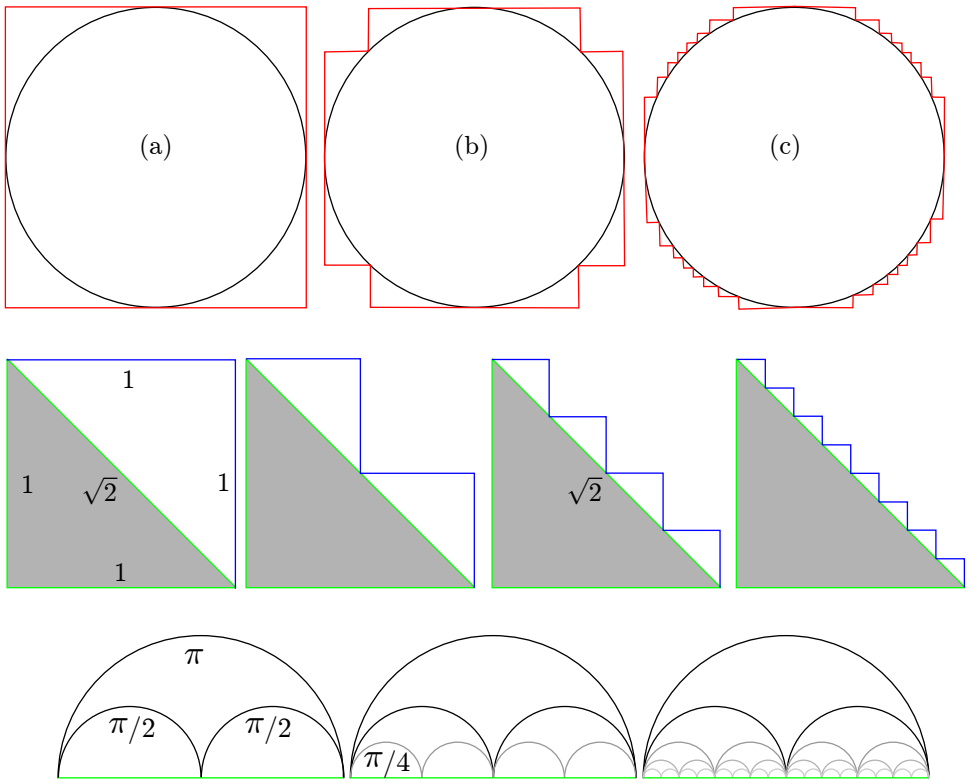


Figure 4.3: A Circle and a Square.

to the circle, so it must be π. What is the explanation? (The other parts of the figure illustrate similar behavior.)

4.5 Set Theory

The examples above illustrate the kind of paradoxes that result when infinity is considered a completed quantity. The conclusion is obvious. We have to be careful when considering processes or situations that involve the infinite. An important example is the unusual equation on Page 133, which involves infinite tetration (repeated exponentiation). Infinity should be considered a potential, a limit, something that cannot be achieved. Yes, there is one area where infinity is considered complete and the resulting paradox is considered acceptable. This area is set theory, an integral part of current mathematics. When first developed and published, in the 1870's and 1880's, by Georg Cantor, some mathematicians, foremost among them Leopold Kronecker, objected to it on the grounds that it (implicitly) assumes that infinite sets can be fully constructed and operated on. Other mathematicians, most notably Hilbert, immediately embraced set

theory and it is easy to see why. Set theory is easy to understand with only a minimal mathematical background. It gives the user the satisfaction of being able to generate and manipulate many different infinite quantities. It is simple and elegant and provides a sense of power.

Because of the popularity of set theory, its story and its history have been told in many books, articles, and websites. This section concentrates on the most important techniques and results of this elegant area of mathematics.

> Cantor's great achievement was to tackle to infinite, to take on the infinite in a way that nobody ever had.
>
> —William Dunham.

Set theory is based on two assumptions. The first assumption is stated explicitly and says that sets can be compared by comparing pairs of elements, one from each set. Imagine two farmers who each own cows. They want to know who owns more cows. They can simply count the number of cows in each set, but what if the farmers cannot count? They can still compare their sets of cows by comparing pairs of cows. Farmer A brings one cow and farmer B brings one of his own cows. Farmer A brings another cow and farmer B does the same. The farmer who runs out of cows first has the fewer cows.



Figure 4.4: Image by MS Copilot, Sept. 2025.

Counting the number of elements of each set is easier and faster, but is possible only if the sets are finite. When dealing with infinite sets, we have to find a rule that will match pairs of elements of the sets. If a rule is found that matches every element of set A with an element of B, such that every element of B is eventually matched, then it is reasonable to claim that the two sets have the same infinite size (the term cardinality is used instead of size). If we can prove that such a rule does not exist, then the two sets, even though both are infinite, can be declared to have different cardinalities.

The second assumption underlying set theory is implicit. It says that we can actually construct complete infinite sets and compare *all* their elements. This is equivalent to talking about infinity as a completed quantity. The following thought experiment may illustrate the difference between potential infinity and completed infinity. Suppose we have balls numbered with the positive integers and we place them one by one into a box. At any point, there will be more balls available to place in the box, infinity hasn't been

reached because it is only a potential. On the other hand, if we imagine a box with *all* the numbered balls already in it, we imagine completed infinity.

Perhaps the simplest example of applying the two assumptions above to actual infinite sets is to compare the set N of positive integers to the set E of positive even integers. Each set is infinite but N includes E, so in a certain sense N is greater that E. However, the simple rule $n \to 2n$ matches each integer n in N with the even integer $2n$ in E. Every integer n is matched by this rule with an even number $2n$ (its double) and every even integer e is matched with an integer $e/2$. If we accept the two assumptions above, then we cannot avoid the conclusion that sets N and E have the same cardinality. They are infinite, but their infinite sizes (cardinalities) are the same.

A similar example compares the set N of positive integers to the set $\mathbb{Z}$ of all the integers. Clearly, $\mathbb{Z}$ includes N, but it is easy to find a rule that matches every element n of N with an element of $\mathbb{Z}$

$$n \to \frac{1+(-1)^n(2n-1)}{4}.$$

The first few matches are $1 \to 0$, $2 \to 1$, $3 \to -1$, $4 \to 2$, and $5 \to -2$, and our rule (again assuming the validity of the assumptions above) proves that $\mathbb{Z}$ and N have the same cardinality.

Cantor used the term countable (or denumerable), to denote the cardinality of N and of any set that can be matched with N. He also used the symbol $\aleph_0$ (pronounced aleph-naught or aleph-sub-zero) to indicate this cardinality. Even more, Cantor considered $\aleph_0$ a new type of number, a type that he termed transfinite.

The next two steps of set theory are really surprising. The first step shows that the set of the rational numbers $\mathbb{Q}$ is countable (its cardinality is $\aleph_0$) and the second step shows that the set of real numbers is not countable, its cardinality is greater than $\aleph_0$ and was termed $\aleph_1$ by Cantor.

A rational number has the form a/b, where a and b are integers and b is nonzero. A rational of the form $a/1$ is an integer, so it is intuitively clear that there are many more rationals than integers. Nevertheless, Cantor found a simple rule to match rationals and integers. We can think of this rule as assigning an integer subscript or tag to each rational number. The rule is illustrated in Figure 4.5 (compare with Figure 4.1). The infinite matrix contains all the rational numbers (many appearing several times), and each is assigned an integer tag by scanning the matrix in a zigzag pattern, as shown in the figure. The result is the infinite sequence 1/1, 2/1, 1/2, 1/3, 2/2, 3/1, 4/1, 3/2, 2/3, 1/4,. . . .

Cantor's proof is not the only one possible. The rationals can be arranged in other ways and matched with the natural numbers. The following table is an example. More and more rationals are included in each group and are arranged such that the sum of the numerator and denominator of the elements of a group is constant. From the way the groups are constructed it is clear that every rational in included in some group. The integer n, for example, first appears as a numerator in group $n+1$ with the denominator 1. All the rationals of the form n/m appear in subsequent groups.

$$\tfrac{0}{1},\quad \tfrac{1}{1},\quad \tfrac{1}{2}\,\tfrac{2}{1},\quad \tfrac{1}{3}\,\tfrac{2}{2}\,\tfrac{3}{1},\quad \tfrac{1}{4}\,\tfrac{2}{3}\,\tfrac{3}{2}\,\tfrac{4}{1},\quad \tfrac{1}{5}\,\tfrac{2}{4}\,\tfrac{3}{3}\,\tfrac{4}{2}\,\tfrac{5}{1}.$$

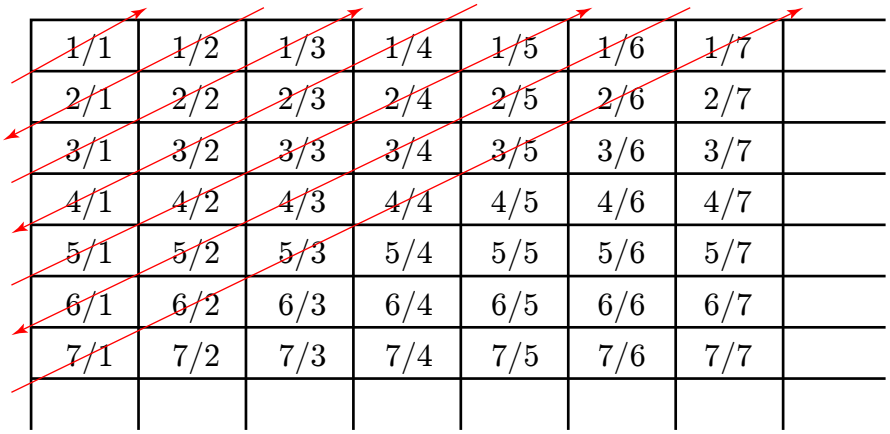


Figure 4.5: Scanning the Rationals in ZigZag.

Now eliminate all the duplicates

$$\frac{0}{1},\ \frac{1}{1},\ \frac{1}{2}\frac{2}{1},\ \frac{1}{3}\ \frac{3}{1},\ \frac{1}{4}\frac{2}{3}\frac{3}{2}\frac{4}{1},\ \frac{1}{5}\qquad\frac{5}{1},$$

close the gaps

$$\frac{0}{1}\ \frac{1}{1}\ \frac{1}{2}\ \frac{2}{1}\ \frac{1}{3}\ \frac{3}{1}\ \frac{1}{4}\ \frac{2}{3}\ \frac{3}{2}\ \frac{4}{1}\ \frac{1}{5}\ \frac{5}{1},$$

and match the natural numbers with the remaining rationals

$$\begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \cdots \\ \frac{0}{1} & \frac{1}{1} & \frac{1}{2} & \frac{2}{1} & \frac{1}{3} & \frac{3}{1} & \frac{1}{4} & \frac{2}{3} & \frac{3}{2} & \frac{4}{1} & \frac{1}{5} & \frac{5}{1}. \end{array}$$

Each integer has a match, and each rational is matched to a different integer. The match is a bijection (a one-to-one mapping).

Here is an even more elegant method, the Stern-Brocot sequence, to generate a complete sequence of the rationals. Start with the recurrence relation: $f(0) = 1$. For $n \geq 1$, $f(2n+1) = f(n)$ and $f(2n) = f(n)+f(n-1)$. This weird relation generates $f(0) = 1$. $f(1) = f(0) = 1$, $f(2) = f(0) + f(1) = 2$, $f(3) = f(1) = 1$, $f(4) = f(1) + f(2) = 3$, $f(5) = f(2) = 2$, $f(6) = f(2)+f(3) = 3$, $f(7) = f(3) = 1$, $f(8) = f(3)+f(4) = 4$, and so on. We end up with the integer sequence $1, 1, 2, 1, 3, 2, 3, 1, 4, 3, 5, 2, 5, 3, 4, 1, 5, 4, 7, \ldots$.

Now convert this sequence to a sequence of rational numbers by (1) start with 1/1, (2) the denominator of each fraction becomes the numerator of the next one, and (3) the numerator of each fraction becomes the next element of the integer sequence above. The result is the interesting Stern-Brocot sequence

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{2}, \frac{2}{3}, \frac{3}{1}, \frac{1}{4}, \frac{4}{3}, \frac{3}{5}, \frac{5}{2}, \frac{2}{5}, \frac{5}{3}, \frac{3}{4}, \frac{4}{1}, \frac{1}{5}, \frac{5}{4}, \frac{4}{7}, \frac{7}{3}, \frac{3}{8}, \frac{8}{5}, \frac{5}{7}, \frac{7}{2}, \frac{2}{7}, \frac{7}{5}, \ldots.$$

This sequence is significant, because reference [Calkin 00] proves that (1) it contains all the rationals, (2) they appear in reduced form, and (3) each appears exactly once.

The surprising conclusion is that even though there seems to be many more rationals than integers, the cardinality of the set of rationals is $\aleph_0$, the same as that of the set N of integers. It is as if including the rationals in N has made no dent in the infinity of N, even though the rationals are dense (see Page 79) and the integers are discrete. This is another peculiar, bizarre, and unexpected feature of the infinite.

It was when Cantor considered the set $\mathbb{R}$ of real numbers, that he received an even bigger surprise. He found it easy to prove that the infinity of $\mathbb{R}$ is, in some precise sense, greater than the infinity of N. In other words, there is no way to match an integer to every real number. Even though there are infinitely many integers, there are "not enough" of them to tag every real number.

The set of real numbers $\mathbb{R}$ consists of the rationals (integers and ratios of integers) and irrationals (numbers such as $\sqrt{2}$, π, and e, that cannot be represented in the form a/b). The proof that there are not enough integers to tag every real number relies on a contradiction and proceeds as follows: Let's restrict ourselves to the short interval $(0, 1)$. Each real number in this interval has the form $0.abcd\ldots$. Let's assume that there is a rule that matches an integer to each real number in this interval. The result of this matching is an infinite list, similar to that of Table 4.6, where each of the real numbers R_i is tagged by (or matched to) an integer i.

$$
\begin{aligned}
1 &\rightarrow R_1 = 0.3572690\ldots = 0.010011101011\ldots \\
2 &\rightarrow R_2 = 0.7096418\ldots = 0.111010010100\ldots \\
3 &\rightarrow R_3 = 0.5381905\ldots = 0.001101001010\ldots \\
4 &\rightarrow R_4 = 0.1033687\ldots = 0.101011001010\ldots \\
5 &\rightarrow R_5 = 0.0036731\ldots = 0.001011010001\ldots \\
6 &\rightarrow R_6 = 0.1958208\ldots = 0.001100001010\ldots
\end{aligned}
$$

Table 4.6: Mapping Real Numbers.

It is now easy to construct a real number S that is not in this list (this will be the contradiction). We simply select the first digit (3) of R_1, the second digit (0) of R_2, and so on, and then modify each digit by, for example, adding 1 to it. The result of this diagonalization process is the real number $S = 0.419481\ldots$. This number is not in Table 4.6 because it differs from any of the real numbers R_i in the table. It differs from R_1 in the first digit following the decimal point. It differs from R_2 in the second digit, and so on. It is easy to construct infinitely many numbers such as S—for example, by selecting the second digit of R_1, the third digit of R_2, etc.—but even showing that one real number is missing from the table is enough to prove that the set of real numbers is uncountable, it represents a higher infinity (i.e., has a larger cardinality) than $\aleph_0$.

◇ **Exercise 4.2:** Come up with a list of real numbers where the diagonalization process does not work and results in a number that is already included in the list.

We now turn to the binary numbers in Table 4.6 in an attempt to (1) understand more about the nature of transfinite numbers, and (2) discover the cardinality of the irrationals $\mathbb{I}$. We assume that the real numbers in the open interval $(0, 1)$ are countable and have binary values as listed in the table. Each binary value consists of $\aleph_0$ bits, so

there can be up to $2^{\aleph_0}$ such values. Thus, there are $2^{\aleph_0}$ real numbers in the interval $(0, 1)$.

◇ **Exercise 4.3:** Prove that this argument applies not just to the interval $(0, 1)$ but to the entire set $\mathbb{R}$ the real numbers, which shows that their cardinality is $2^{\aleph_0}$.

Thus, the cardinality of both the reals $\mathbb{R}$ and the irrationals $\mathbb{I}$ is $2^{\aleph_0}$. However, applying the same argument to the decimal numbers of Table 4.6 suggests that the cardinality of both the reals and the irrationals is also $10^{\aleph_0}$, which forces us to conclude that $2^{\aleph_0}$ and $10^{\aleph_0}$ are the same cardinalities, and that the transfinite numbers are different from the real numbers. They obey different rules and should perhaps be called transfinites instead of transfinite numbers.

Cantor denoted the cardinality of the reals by $\aleph_1$. After years of searching, he was unable to find a set whose cardinality is between $\aleph_0$ and $\aleph_1$, but was also unable to prove that such sets do not exist. The question of whether such sets exist became known as the continuum hypothesis, and was not settled until the 1960s.

The quantities $\aleph_0$ and $\aleph_1$ became known as transfinite numbers or transfinites, and it is easy to construct infinitely many of them. The infinity of the set of transfinites is itself countable, since each transfinite number is tagged by an integer.

The results so easily achieved by set theory are surprising, unintuitive, and impressive, but they always raise the question of what is all this good for? One surprising result is the $\aleph_1$ cardinality of the irrationals. After all, very few irrational numbers are known, while we feel that we "know" all the rationals. Another surprise was Cantor discovery that the cardinality of the transcendental numbers is $\aleph_1$, while the cardinality of the algebraic numbers (which include both rationals and irrationals) is only $\aleph_0$. Any real number is either algebraic or transcendental. A number is algebraic if it is the solution to a polynomial equation with integer coefficients. Any real number that is not such a solution is termed transcendental because it transcends algebra. Very few transcendental numbers are known (among them are e and π), so mathematicians tended to assume that they are rare, but Cantor found a rule that matched an integer to each algebraic number, thereby showing that the cardinality of the algebraic numbers is only $\aleph_0$ and implying that the cardinality of the transcendentals is $\aleph_1$. This existence proof (Cantor did not discover any transcendentals, he only proved that they exist) is a surprising result that is equivalent to the following. If you select a real number at random among *all* the real numbers, it will be transcendental with probability 1!

> The algebraic numbers are spotted over the plane like the stars against a black sky; the dense blackness is the firmament of the transcendentals.
>
> —E. T. Bell.

◇ **Exercise 4.4:** Try to prove that the set of algebraic numbers is countable (i.e., has cardinality $\aleph_0$). As a last resort, locate such a proof and try to understand it.

Once we accept the existence of the transfinite numbers $\aleph_0$ and $\aleph_1$, a natural question is whether larger transfinite numbers exist. The elegant and surprising answer of set theory is yes. There is a (countable) infinite sequence of transfinite numbers $\aleph_i$, for $i = 0, 1, 2, \ldots$, and in principle it is easy to construct a set whose cardinality is $\aleph_i$ for any given i. Like many ideas and concepts in set theory, the process of constructing bigger

and bigger infinite sets is conceptually simple. Given a set of cardinality $\aleph_{i-1}$, the set of all its subsets (termed the power set of S) has cardinality $\aleph_i$.

This construction is easy to illustrate with finite sets. Given the set $S = \{A, B, C\}$, the set of all the subsets of S is $\{\{\emptyset\}, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, S\}$ (notice that the empty set $\emptyset$ and S itself are considered subsets of S). If the original S has n elements, then the set of all its subsets has 2^n elements.

Given that the set $\mathbb{R}$ of real numbers has cardinality $\aleph_1$, its power set $T = P[R]$ will have cardinality $\aleph_2$. T is simply the set of all subsets of real numbers. Similarly, set $U = P[T]$ (the set of all subsets of elements of T) will have cardinality $\aleph_3$, and so on.

Set theory is based on the two assumptions above and its main conclusions are (1) there are infinitely many infinities and (2) there is no largest infinity. Deep, unexpected results whose validity depends on the truth of the two basic assumptions.

The discussion here is limited to the most important results of set theory. Cantor went beyond these results and proved more interesting theorems about infinite sets, especially those whose elements can be ordered. In his lifetime there was strong opposition to his work from important mathematicians such as Kronecker and Gauss, but today set theory is regarded as a bona fide part of mathematics and transfinite numbers are considered almost as ordinary as the integers.

> Some of you may have met mathematicians and wondered how they got that way.
> —Tom Lehrer.

4.5.1 The Cardinality of Power Sets

It is easy to compare the sizes of finite sets, and it is similarly easy to see that a finite set is always smaller than its power set. When it comes to the cardinality of infinite sets, however, things are not so simple and size relations have to be proved rigorously. Already Cantor himself proved that the cardinality of a power set is always greater than that of its parent set, and this subsection presents a simple proof of this fact, a poof that leads to a paradox.

Imagine a set S that may be finite, infinite, countable or uncountable. We denote its power set by P. Each object s of S belongs to many subsets that are elements of P. The proof proceeds in two Steps.

Step 1. We construct a map from each element s of S to a distinct element p of P. This will prove that the cardinality of S is less than or equal the cardinality of P. Some of the elements of P are single-element subsets of S. Each of those unit sets or singletons consists of just one element of S, and we map that element to this unit set. This one-to-one mapping of S to part of P shows that the cardinality of S is at least less than the cardinality of P (but it may conceivably be equal to it)

Step 2. We show that S and P have different cardinalities. This will imply that the cardinality of S must be less than the cardinality of P. The cardinalities would be the same only if there is a one-to-one mapping of S to P, so our task is to show that there can be no such mapping. We assume that there is one, and then end with a contradiction.

We examine the elements of P. Each is a subset of S and each will now be assigned a representative (or rep for short) that is an element of S. Also, each s in S will become a rep of one distinct p of P. If the rep of some p is also an element of p (recall that each

p is a subset of S), then it is referred to as an internal rep. Otherwise, it is an external rep. For example, the empty set $\emptyset$ is an element of P, but since it is empty, its rep would have to be external. This example also shows that there must be at least one external rep. As another example, consider that S, being its own subset, must be an element of P, so its rep must be internal, thereby proving that there must always be at least one internal rep.

Once this is grasped, it should be clear that the set X of all external reps, being a subset of S, is itself an element of P, and therefore must have its own rep x. Is $x \in X$? Clearly, x cannot be a member of X, because only external reps are members of X. But since x is not a member of X, it is an external rep and therefore must belong to X. We have arrived at a contradiction, and must agree that x simply cannot exist. On the other hand, x must exist if there is a one-to-one mapping of S to P, which proves that there is no such mapping, and the cardinality of P is strictly greater than the cardinality of S.

We are now ready for the Cantor paradox. This paradox involves a real monster, the set L of all sets. Imagine such a set! its elements are sets, and they constitute all the sets there are. L is the biggest set of all the sets whose elements are sets; a true leviathan. However, its power set is also a set of sets and its cardinality must be strictly greater than the cardinality of L. Thus, L is both the biggest and not the biggest of all the sets whose elements are sets; a paradox.

Attempts to resolve the Cantor paradox try to refer to L as an illegitimate set, a pathological set, or even as a proof that set theory is entirely wrong and should be ignored.

■ Paradox of Omniscience. Some people believe that God knows everything, but this belief opens the way to a paradox, one that is based on Cantor's paradox. In order to focus the problem, we suppose that there is a deity that knows, not everything, but only all truths. There must therefore be a set T (finite or infinite) that contains all the possible truths. According to Cantor, the power set P of T must be larger than T, and we use this fact to prove that T cannot contain all truths.

Suppose that a is one of the truths in T. It would therefore belong to some of the subsets p that are elements of P, such as (a, b, c) or (a, g, y), but not to other subsets, such as $\emptyset$ or (c, d, e). This implies that for each element p of P, one of the following statements "a is in p" or "a is not in p," is true. We refer to these statements as truths of type X. There is a truth of type X for each element p, but there are more elements p in P than elements t in T, which shows that T cannot be a set of all truths.

> If everybody contemplates the infinite instead of fixing the drains, many of us will die of cholera.
>
> —John Rich.

4.6 Physical Infinity

Consider the following experiment. Two parallel mirrors are separated by a certain distance. A beam of light is sent to one mirror and is reflected to the other mirror. If the beam is exactly perpendicular to the mirrors, it should bounce back and forth between them an infinite number of times. However, the physical world is constructed such that this does not happen. A real mirror does not reflect 100% of the light, and some light is absorbed by the air between the mirrors. Thus, after many reflections, only one photon of light will remain and it will eventually be absorbed, ending this process. In addition, the speed of light is finite, which prevents it from traveling an infinite distance even under otherwise ideal conditions.

In general, our world is quantized. Matter consists of atoms and elementary particles, while electromagnetic radiation consists of photons. It is impossible to break a piece of matter up into smaller and smaller components forever, because an elementary part is eventually reached. Similarly, it is impossible to isolate smaller and smaller units of energy, because an indivisible photon is eventually reached.

The Heisenberg uncertainty principle also limits what can exist. Current science acknowledges the Planck scale, a measure that consists of the Planck energy/mass (2.2×10^{-8} Kg), Planck length (1.6×10^{-35} m), and Planck time (5.4×10^{-44} s). The Planck length is related to Planck energy/mass by the uncertainty principle. At this minute scale, our intuitive concepts of time and distance break down and quantum indeterminacy dominates. The Planck time is the time light takes to travel, in a vacuum, a distance of one Planck length. According to current physical theory, the Planck time is the smallest time measurement that can in principle be possible.

Thus, it seems that the only thing in the physical world that may turn out to be infinite is the universe itself. It may be infinite in size and may exist forever, but currently it is impossible to verify or disprove any of this. It may also exist forever, but even this is currently unknown. The old steady-state theory assumed that the universe is infinitely large, has always existed and will always exist. The galaxies move away from each other, but new atoms are created out of nothing in the space between galaxies and slowly coalesce to form new galaxies. This was an elegant theory that gracefully embraced the infinite both in space and time, but in the 1960s it was discredited in favor of the big-bang theory.

A more practical question is whether there may be infinite quantities, such as infinite density, speed, or temperature somewhere in the universe, currently or in the finite future. The big bang theory (a current cosmological favorite) assumes that the universe started with an explosion, but current physics is unable to probe the details of this explosion at its very beginning (within the Planck time of the beginning).

Black holes are very much in fashion nowadays, and they may be the only candidate for physical infinity. A black hole starts as a star that is at least three times more massive than our sun. When its nuclear fuel runs out, such a star starts collapsing on itself. Because of its large mass, the gravity it generates is so strong that it collapses the star to a mathematical, dimensionless point (a singularity) of infinite density. The gravity of a black hole is so strong, that anything (particles or photons) passing closer to it than a certain distance (its event horizon) is attracted to it, cannot escape, and must fall into the singularity. As a result, it is impossible to see inside the event horizon

and actually observe the singularity, which is the reason for the name black hole.

Because of this feature of black holes, Roger Penrose has suggested that there may be a science principle or a law of nature that may be termed cosmic censorship, and that makes it impossible to observe any physical singularity. If this is true, then any singularity will be hidden by some event surface (no naked singularities), but it is currently not known whether nature obeys such a principle. In the past natural philosophers claimed that nature abhors vacuum. Today, we may claim that nature hides all naked singularities.

Can there be infinite temperature? There are two reasons to believe that this is impossible: (1) In order to heat something to a high temperature, a heat source of a higher temperature is needed. (2) Heat is a form of energy. It is the kinetic energy (energy of motion) of atoms and molecules. Thus, temperature is an indication of the speed of these particles, and infinite temperature requires infinite speed, which is not allowed by special relativity. A particle moving at the speed of light has infinite energy and thus corresponds to infinite temperature, but there may not be an infinite amount of energy in the entire universe.

Another limit to how hot it can get is the relation between temperature and emitted radiation. Car owners know from experience how hot a car can get on a sunny day. However, a car, like any other object, does not get hotter without a limit, because as its temperature increases, it also emits more radiation, in the form of infrared photons. In fact, any object at a temperature higher than absolute zero emits electromagnetic radiation whose wavelength gets shorter (and its frequency longer) as the temperature climbs.

At the Planck temperature, 1.41×10^{32} Kelvin, the wavelength of the emitted radiation gets down to approximately 2×10^{-35} meter. This is very close to the notorious Planck length, the shortest possible distance in our physical universe according to the current quantum mechanics. Thus, the Planck temperature is the current accepted limit of high temperatures.

Infinite speed is forbidden by special relativity, but this applies only to material objects or to entities that carry information. So what can move at infinite speed? Here are two examples:

Example: Given a pair of scissors we fully open them, and then close them in one second. When the two scissor blades are being closed, the point between them moves toward the tips of the blades. We also know from experience that this point accelerates as it moves. Now imagine a thought experiment involving a pair of scissors with extremely long blades, perhaps hundreds of kilometers long. We open them and close them in one second. The point between the blades now has to move the entire length of the blades in one second. If the blades are long enough, the speed of the point at the end of its travel may be infinite.

Example: Cut a branch off a tree and stick it in the ground, so it stands vertically. As the sun moves overhead, the branch casts a shadow. In late afternoon, the shadow lengthens. In the last second before the sun sets below the horizon, the shadow has to move an infinite distance, so in principle, its speed goes all the way to infinity. In practice, the shadow gets so weak that we completely lose it.

Until 1915, space was considered static; an unchanging stage on which actions take place. The publication of general relativity in 1915 changed this point of view. Space is

now seen as flexible fabric whose shape depends on the masses within it. A large mass bends space in a way that we cannot visualize but can compute.

The discovery of the expansion of the universe, in the 1920s, did much to extend our understanding of the universe. Current cosmology says that the expansion of the universe may have three results: (1) The expansion may stop and turn into a contraction. (2) It may go on forever and separate the galaxies more and more. (3) It may continue forever, but slow down gradually, never actually coming to a stop.

> No continuous thing is divisible into things without parts.
> —Aristotle.

4.7 Infinitesimals and the Calculus

An infinitesimal number is a non-zero quantity that is closer to 0 than any non-zero real number is. The word infinitesimal comes from a 17th-century Modern Latin coinage infinitesimus, which originally referred to the "infinity-eth" item in a sequence. (From Wikipedia.)

Infinitesimals play an important role in the differential calculus, which makes them part of mathematics. In some sense, an infinitesimal is the opposite of infinity, which is why this section is included in the chapter on the infinite.

The ancients were already interested in infinity. This is obvious from Zeno's paradoxes and from the writings of Democritus about the possible atomic nature of objects.

As far as we know, the modern interest and serious mathematical research in infinitely small quantities started in the 1600's by the Italian monk and mathematician Bonaventura Cavalieri. Among his many letters to many correspondents, one stands in particular. A 1621 letter to Galileo that shows how Cavalieri introduced infinity into geometry by thinking about the area of a geometric objects as an infinite collection of infinitely thin lines. Similarly, he considered the volume of geometric objects as an infinite collection of infinitely two-dimensional thin slices. Reference [Alexander 14] has brought Cavalieri and his methods to my attention, and the following paragraph summarizes this historic letter.

Given a two-dimensional plane figure, we start by drawing a straight line inside it, and then draw "all" the possible lines that are parallel to this first line and are inside the figure. Is it permissible to equate the plane figure with this set of "all" the lines? Similarly, given a three-dimensional solid, start by drawing a thin plane contained inside the object, and then draw all the possible planes that are parallel to the first plane and are inside the object. Is it permissible to equate the object itself with the set of "all" those planes? In general, does it make sense to compare "all" the lines of two figures? This kind of operation implies comparing one infinity with another, an operation that mathematicians have traditionally avoided, since it can easily lead to contradictions and non-sensical results.

Galileo's answer would be of interest to us, but unfortunately was not preserved and remains unknown. However, the answer, and perhaps subsequent answers by Galileo, must have been encouraging, because Cavalieri continued to pursue those ideas and started referring to his thin lines and slices as indivisibles. He also published several

proofs of geometry based on them. A typical such proof shows that two plane figure have the same area because they "consist" of the same number of identical indivisible lines placed side by side.

While correct, his proofs are intuitive and not rigorous. Also, such proofs are easier to establish using the classical methods of Euclid, but Cavalieri wanted to show that there is something in his ideas of the indivisibles. It is obvious to us, as doubtless it was obvious to him, that there is a problem with indivisibles, because there must be infinitely many of them. If each line or plane is infinitely thin, then the sum of their widths must be zero, and if their widths are nonzero, then their total sum must be infinite.

Today, Cavalieri's ideas are known as Cavalieri's principle, a proposition which has two parts as follows:

- The two-dimensional case: Suppose that two regions in a plane are included between two parallel lines in that plane. If every line parallel to these two lines intersects both regions in line segments of equal length, then the two regions have equal areas.

- The three-dimensional case: If, in two solids of equal altitude, the sections made by planes parallel to and at the same distance from their respective bases are always equal, then the volumes of the two solids are equal.

Even though his writing was dense and opaque, Cavalieri's ideas spread through Europe, because he wrote widely and often and he corresponded with other mathematicians such as Marin Mersenne, Vincenzo Viviani, and Evangelista Torricelli. The latter especially is responsible for furthering and promoting Cavalieri's ideas.

> Few, if any, since Archimedes, have delved as far and as deep into the science of geometry.
>
> —Galileo in a letter to Cesare Marsili, March 10, 1629, about Cavalieri.

The next step in our understanding of infinitesimals and infinity was made by Evangelista Torricelli, a contemporary of Cavalieri's. Today, Torricelli is mostly known as the inventor, in 1643, of the barometer, but in his lifetime he was known for several important works both in mathematics and in physics. The following paradox, Figure 4.7(a), illustrates his approach to Cavalieri's indivisibles

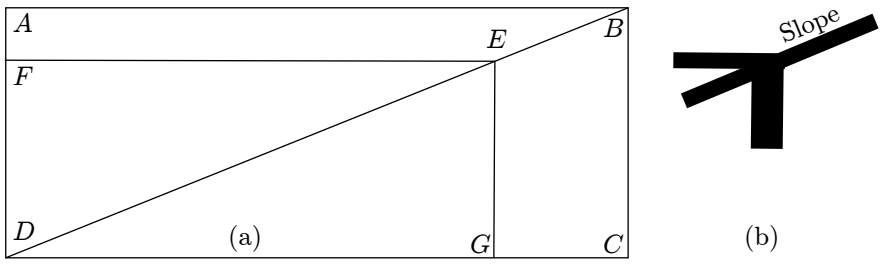


Figure 4.7: Torricelli's Paradox.

Given the rectangle (which can also be a parallelogram) $ABCD$, draw a diagonal DB which partitions it into two equal triangles. Draw a horizontal segment FE at an

arbitrary height, and the vertical segment that connects point E to the bottom face at a point G. Notice that EF is longer than EG, which implies that all the horizontal lines parallel to EF in the upper triangle are longer than their corresponding vertical lines in the lower triangle. If we assume that the set of ALL the lines in a triangle determine its area, we must conclude that the upper triangle is greater than the lower triangle; a paradox.

When early mathematicians, among then ancient Greek geometers, were faced with such contradictions, they either ignored them or banned them. When Cavalieri met with similar contradictions caused by his method of drawing infinitely many parallel indivisibles, he tried to introduce artificial rules to eliminate or suppress his difficulties. Torricelli, on the other hand, tried to get to the bottom of those paradoxes, and the way he resolved this paradox is original and important because it (1) introduced the concept of a slope, and (2) opened the way for future mathematicians to extend his ideas and come up with the modern concept of a limit.

Torricelli started by saying "let's forget for a moment about infinitely thin lines, and assume the our horizontal and vertical lines have finite widths." The consequence of this assumption is obvious and immediate, and is illustrated in Figure 4.7(b). Because $ABCD$ is a rectangle, sides AB and BC have different lengths, which causes the diagonal DB to have a slope of less than $45°$. This implies that the vertical lines are now thicker than the horizontal lines, which compensates for the smaller number of these lines and resolves the paradox.

Ingenious, but what happens when we now gradually shrink the widths of the line segments back to become indivisibles? At the "last moment," when the number of lines becomes infinite, the paradox pops up again. This indicates that there is a fundamental difference between the finite and the infinite. When dealing with indivisibles, we must assign them different "widths" even though they are supposed to be infinitely thin. Geometrically, all four sides $ABCD$ of the rectangle have the same width, but when we deal with them as indivisibles, even the sides themselves, such as AB and BC must be thought of as having different widths. Also, the ratio of their widths (which is the slope) is the inverse of the ratio of their lengths.

We cannot simply draw conclusions by extending concepts and results from the realm of the finite to that of the infinite. Perhaps the best way to deal with infinities is to start with the finite and approach the infinite gradually. This realization is what led later mathematicians to the important concept of the limit.

The concept of a slope, illustrated in this paradox, is simple and the slope in this example is constant, because the diagonal is straight. Once mathematicians realized the important of the slope in dealing with indivisibles, they extended it to curves, where the slope is variable. Today, we are familiar with the calculus and the concepts of variable slope and rate of change, but it was the work of these early Italian mathematicians that opened the way for Newton, Leibniz, and others to introduce and develop the calculus.

Lacking scientific journals, 15th century mathematicians communicated by letters, naturally in Latin, as well as by writing textbooks. Cavalieri corresponded with French and Italians mathematicians, and Torricelli, similarly, corresponded with Gilles Personne de Roberval, Pierre de Fermat, and Marin Mersenne. The latter corresponded with, among others, John Wallis and Isaac Barrow.

Such communications proved crucial for the dissemination of new ideas, because, as

the next few paragraphs explain, Cavalieri, Torricelli, Galileo, and others in Italy were silenced by the Catholic Church and had to stop writing, communicating, and teaching their ideas about indivisibles and other topics. Fortunately, the political and religious climate in central and Northern Europe was more tolerant and conducive to scientific ideas, which permitted progress and allowed Newton and Leibniz to develop the calculus in the 1660's.

Pope Julius II died in 1513 and Giovanni de Medici became his successor as pope Leo X. Even though clever and extremely educated, Leo proved more interested in having a good time than in administering the Church. He borrowed and spent money without circumspection, partying and supporting the arts, and he extended the practice of selling indulgences in order to balance the Church's budget.

In 1510, the lowly priest Martin Luther was sent to Rome together with one of his Augustinian brothers, and what he saw there appalled, disillusioned, and confused him. He realized that the Pope, his Cardinals, and the entire Vatican hierarchy were corrupt. He later wrote "if there is Hell, then Rome is built over it." Back in Wittenberg, he collected his grievances against the Church and wrote his historically important 95 theses, published in 1517. This started the reformation which succeeded within a few dozen years to considerably reduce the power and influence of the Catholic Church. The story of the reformation is, of course, well known and widely documented, but the part that is relevant to our story of the indivisibles and the treatment of infinities is less known.

There is one rule which I would recommend to your attention in preference to all others. Rise early in the morning. This will not only contribute to your health, but will enable you to arrange and expedite the business of the day; and as there are various duties incident to.

—Lorenzo de' Medici to his son Giovanni.

God has given us the papacy, let us enjoy it

—Leo X (possibly).

This part started inconspicuously on September 27, 1540, when Pope Paul III approved a petition from a group of 10 priests to form a religious order dedicated to serving the Pope and the Church. Once approved, its 10 founders named the new order the Society of Jesus, and became known as Jesuits. The new order was mainly the brainchild of Ignatius of Loyola, an unknown Spanish priest. The story of the Jesuits is also well known and well documented. Suffice it to say that this order, which grew quickly, saved the Catholic Church and stemmed the growth of the protestant revolution.

A unity among many cannot be maintained without order, nor order without the due bond of obedience between inferiors and superiors.

—Ignatius of Loyola.

Unfortunately for the development of science and mathematics, this success of the Jesuits also slowed down progress for several centuries, as the Jesuits quickly formed the censorship department known as the Revisors General, which had the authority to approve or condemn ideas and works in all areas. On August 10, 1632, the revisors had issued their decision on the study of indivisibles as follows "We consider this proposition

to be not only repugnant to the common doctrine of Aristotle, but that it is by itself improbable, and ... is disapproved and forbidden in our Society."

Hence, the next chapter in the story of the indivisibles starts in England, far from Italy and the Church, and the next character in this story is John Wallis—a clergyman, a contemporary of Newton, and a well-known mathematician and cryptographer—on whom history confers partial credit for the development of the infinitesimal calculus (no longer indivisibles).

> Mon ami, a clue of two feet long is every bit as valuable as one measuring two millimeters! But it is the romantic idea that all important clues must be infinitesimal.
>
> —Agatha Christie, *Murder on the Links*, 1923.

In his early career, Wallis made a name for himself as both a clergyman and a clever cryptanalyst, not as a mathematician, which is why he, as well as his friends and acquaintances, were surprised when in 1649 he was suddenly appointed Savilian Professor of Geometry in Oxford. The 1640's were a turbulent time in British history, with the civil war and the overthrow of the monarchy, which suggests that his appointment was based largely on political grounds. Nonetheless, Wallis more than justified his new job by the work he did through the 54 years he held it.

Specifically, Wallis made significant contributions to trigonometry, calculus, geometry, and infinite series. He also coined the term "continued fraction" and is considered the originator of the important concept of the number line. On the other hand, his approach to mathematics was not along the lines of the Greek tradition of exact proofs and absolute results. He preferred to proceed through trial and error, obtaining results that were extremely probable but not irrefutably certain; results that would be validated not through logic and pure reason, but by consensus. In this way, he became known as a prolific worker who came up with many new and effective results that were not always perfect.

His approach to the indivisible was also practical rather than abstract. He knew about the work of Cavalieri and his Italian successors, but Wallis replaced Cavalieri's infinitely thin lines with parallelograms of finite width, a width that he denoted by $1/\infty$. He shrugged off any objections and criticism from contemporaries such as Thomas Hobbes and Pierre Fermat, ignored the contradiction of his infinitely many parallelograms adding to an infinite width, and felt that mathematical rigor and complex, precise proofs were superfluous. A typical example of his style is demonstrated in an early work where he computed the area of a triangle. This work contains the weird expression $\frac{1}{\infty}A \times \frac{\infty}{2}B = \frac{1}{2}AB$, where he has the cheek to mix up real numbers with transfinites, as well as the impertinence to reduce the term $\frac{\infty}{\infty}$ to 1, an unforgivable sin in current mathematics.

The next step in the history of the infinitesimal was taken independently by Newton and Leibniz. The former developed this field, which he referred to as fluxions, in order to put his laws of motion and universal gravitation on a firm mathematical basis. The latter spent much effort developing better notation and was also the first to state and list the rules of calculus.

However, after these giants did their work and laid out the foundations of calculus, much room was left for extensions, improvements, and for building a rigorous mathematical basis for the concepts of function, limit, and rate of change. The important

concept of limit was especially treated and defined by Bernard Bolzano, Augustin-Louis Cauchy, and Karl Weierstrass. The latter also introduced the notation lim and $\lim_{x\to 0}$.

As a result of their efforts we have the modern definition of a limit. The statement: A function $f(x)$ has a limit L when x approaches p, means that for every real number $\epsilon > 0$, there exists a real $\delta > 0$ such that $0 < |x - p| < \delta$ implies that $|f(x) - L| < \epsilon$.

Finally, a word about the calculus of variations. This practical part of the differential calculus uses variations, small changes in functions, to find maxima and minima of functions. An interesting example is the well-known problem of the lifeguard and drowning person, which is illustrated in Figure 4.8. The lifeguard tries to reach the drowning person as fast as possible, and three possible routes are indicated in the figure. Route 1 minimizes the distance the lifeguard has to run on land, while route 3 minimizes the swimming distance. Since swimming is slower than running, the latter route makes more sense, but may not be the best one. Route 3 may be better, but the solution must determine the exact point x where the lifeguard should enter the water.

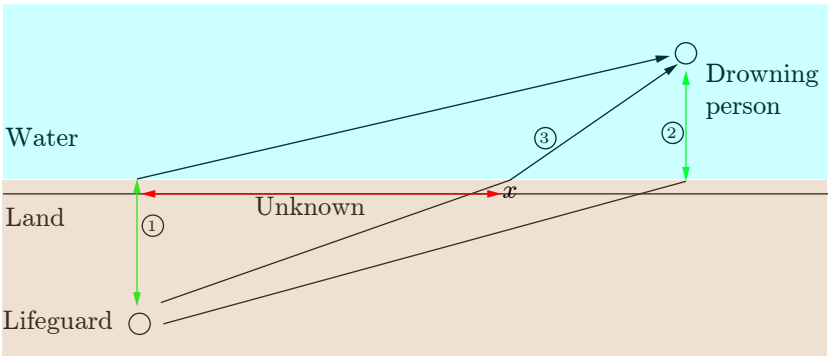


Figure 4.8: Lifeguard and Customer.

In order to solve the problem, the distances (in green) must be known as well as the speeds of the lifeguard in water and on land. A simple analysis produces the total time as a function of these quantities and of the unknown distance x. The function is then differentiated with respect to x, and the derivative is equated to zero, which provides an equation for the unknown x. This process, the culmination of hundreds of years of research, is currently taught in high schools.

4.8 Restrictions in Mathematics

Mathematics is a vast field of knowledge. It is a creation of the human mind, but it also applies to and helps explain the real world around us. It is based on logic and allows for no personal opinions or beliefs. It is also infinite and may continue to provide us with infinitely many concepts, theories, and interesting problems. It is also rigorous. Mathematical definitions must be precise and should not result in contradictions, absurdities, or different interpretations. This is why mathematics has its limits and must impose certain restrictions on what can and cannot be done. What is allowed and what is invalid. This section describes a few such restrictions that may confuse and puzzle students of mathematics. Most of these restrictions are the result of differences between real numbers and complex numbers, and they make it clear that any computations involving complex numbers must be done carefully and reviewed before they can be considered correct.

- Dividing by zero. As we divide a real number a by smaller and smaller numbers, the quotient gets bigger, which is why students of mathematics tend to extend this fact intuitively and claim that division by zero yields infinity. This claim may make sense as long as we realize that infinity is a concept, not an actual number, and cannot be operated on. Hence, students are taught that division by zero is prohibited, but this rule is a result of experiments that show what may happen if the rule is not obeyed.

  Division is also considered the inverse of multiplication, and we expect a number to go back to itself when passed through an arithmetic operation and its inverse. However, when dividing a by zero to obtain infinity, and then multiplying this infinity by zero, the result is zero, not a.

  An old, well-known example of an absurdity caused by division by zero is

$$\begin{aligned} x &= 1, \\ x &= y, \\ x^2 &= xy, \\ x^2 - y^2 &= xy - y^2, \\ (x+y)(x-y) &= y(x-y), \\ x+y &= y, \end{aligned}$$

  from which we can conclude that either $x = 0$ and $y = 1$, implying $0 = 1$, or that $2y = y$, which implies $2 = 1$. The reason for the absurd results is the last step, where we divide by $(x - y)$ which is zero.

- The expression 0^0 is considered invalid by certain mathematicians. Here are several reasons for and against this:

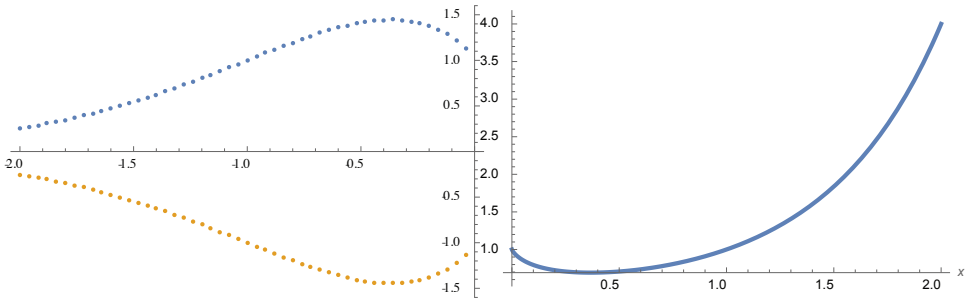
  1. For any real $a \neq 0$, we agree that $a^0 = 1$, because $a^0 = a^{1-1} = a^1 \cdot a^{-1} = a/a = 1$. However, applying this idea to $a = 0$ result in $a^0 = 0/0$, an expression which we already know leads to wrong results.

  2. The exponentiation operator is defined as repeated multiplication. Thus, for example, $a^4 = a \cdot a \cdot a \cdot a$, which implies that $0^4 = 0 \cdot 0 \cdot 0 \cdot 0 = 0$. On the other hand

we have $a^0 = 1$, which implies that 0^0 should also be 1. This ambiguity suggests that 0^0 should be declared invalid.

3. In order to get an idea of a possible value for 0^0, it makes sense to check the behavior of the function $f(x) = x^x$ in the vicinity of $x = 0$, and to approach this value twice, from the positive and from the negative sides. Using mathematical software, it is clear that as we move down from $x = 1$ to $x = 0$, $f(x)$ first drops from 1 and becomes zero at $x = 1/e \approx 0.367879$, and then rises steadily and approaches 1 as x gets closer to zero.

For negative x, however, the value of $f(x)$ is almost always complex. Reference [Meyerson 96] shows that $f(x)$ returns real values for negative x only when x is of the form $-p/q$, where p and q are positive integers and q is odd. Figure 4.9 illustrates this by plotting x^x at many points in the negative region, and as a continuous curve for positive x.



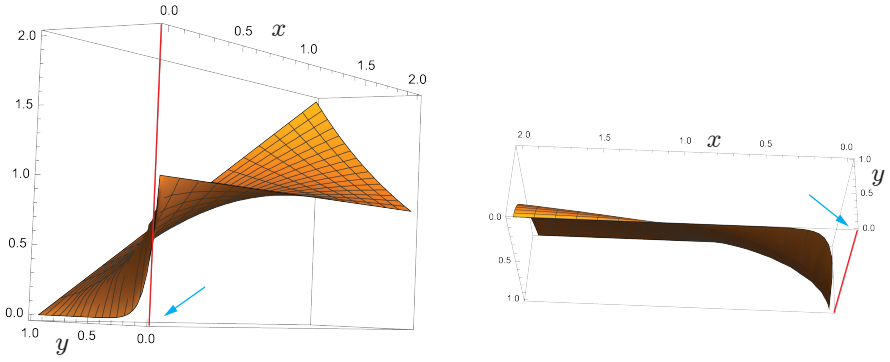

```
(*Plot x^x for positive and negative values of x*)
Plot[x^x,{x,0,2}, AxesLabel->Automatic]
tp=Table[{x,Abs[x]^x},{x,-1/25,-50/25,-1/25}];
tn=Table[{x,-Abs[x]^x},{x,-1/25,-50/25,-1/25}];
ListPlot[{tp,tn}]
```

Figure 4.9: The Behavior of x^x Around Zero.

◇ **Exercise 4.5:** Find the x value where x^x has a minimum.

4. Consider the two-variable function $f(u, v) = u^v$. For a constant u, this function is exponential as, for example, in $f(2, v) = 2^v$, but for a constant v, $f(u, v)$ is a power function, as in $f(u, 2) = u^2$. We now consider the limits $\lim_{u\to 0} f(u, v)$ and $\lim_{v\to 0} f(u, v)$. The former is the limit of an exponential function at zero, while the latter is the limit of a power function at zero, and Figure 4.10 illustrates that these limits are different. The edges where $x = y$ are indicated in red, and the origin is pointed to by an arrow.

■ The square root of a negative number must be paid special attention. It should be treated as the square root of a product as follows $\sqrt{-a} = \sqrt{(-1)\cdot a} = \sqrt{-1}\sqrt{a} = i\sqrt{a}$. One example is the product $\sqrt{-81}\sqrt{-4}$ which equals $i\sqrt{81}\cdot i\sqrt{4} = -9\cdot 2 = -18$. Another example is the equation $\sqrt{x} + \sqrt{-x} = a$. If we simply multiply both sides by $\sqrt{x}$, we end up with $x + (-x) = a\sqrt{x}$ or $x = 0$, which is wrong. The solution is easy to figure out, and it must be imaginary.




```
Plot3D[x^y, {x, 0, 2}, {y, 0, 1}, BoxRatios -> {2, .5, 1},
 AxesLabel -> Automatic]
Plot3D[y^x, {x, 0, 2}, {y, 0, 1}, BoxRatios -> {2, .5, 1},
 AxesLabel -> Automatic]
```

Figure 4.10: Two Surfaces to Illustrate x^y.

◇ **Exercise 4.6:** Solve the simple equation $(-2)^x = 2$ twice, using different methods and paying attention to the negative sign.

■ Replacing $\sqrt{a}/\sqrt{b}$ with $\sqrt{a/b}$ is valid only if a is non-negative and b is positive. Trying to do such a thing when a or b are negative leads to wrong results, as is demonstrated by Exercise 2.44. This is therefore another rule or restriction in the use of mathematics and arithmetic operations.

■ We intuitively feel that the expressions $\sqrt[n]{a}$ and $a^{1/n}$ are identical, but this is true only for nonnegative numbers. There is a difference between a root of a negative number a and the exponents of a. When we type `CubeRoot[-1]` in Mathematica, it returns -1, but when we try `(-1.)^(1/3)` the result is the complex number `0.5+0.866025I`, which is definitely not negative 1. This is explained on Page 111 and is also discussed in reference [MindYourDecisions 23].

■ The square root of a positive number must be positive. A negative square root of a positive number may lead to wrong and absurd results. The simple equation $x^2 = 4$ has the solution $x = 2$, because $2^2 = 4$. But the same equation also has the solution $x = -2$, because $(-2)^2 = 4$. We therefore tend to assume that the $\sqrt{4}$ has the two values ± 2, but this leads to contradictions as shown, for example, in reference [MindYourDecisions 24], where this assumption is used to prove that $2 + 2 = 5$.

■ Fractions can be reduced as, for example, in $10/6 = 5/3$, but it is risky to reduce an exponent in the same way, such as going from $(a + b)^2 = (c + d)^2$ to $a + b = c + d$.

Here is an example where the wrong step is indicated by a *:

$$\begin{aligned} 1 &= 1, \\ 13 - 12 &= 25 - 24, \\ 4 + 9 - 12 &= 16 + 9 - 24, \\ 2^2 + 3^2 - 2 \cdot 2 \cdot 3 &= 4^2 + 3^2 - 2 \cdot 4 \cdot 3, \\ (2-3)^2 &= (4-3)^2, \\ 2 - 3 &\stackrel{*}{=} 4 - 3, \\ -1 &= 1. \end{aligned}$$

■ The simple power rule $(e^x)^y = e^{xy}$ is always true for real numbers x and y, but may sometimes lead to wrong results for complex numbers. The correct rule (correct in the sense that it produces correct results not just sometimes but always) for complex exponentiation is

$$Z^W = e^{W \ln Z}. \tag{4.7}$$

As an example, consider the complex expression $(e^{2i\pi+1})^{2i\pi+1}$. Using mathematical software, which knows to apply Equation (4.7), this is shown below to equal the (real) Euler constant e, but when we write this expression as $e^{(2i\pi+1)(2i\pi+1)}$, the same software computes it as `1.94552*10^-17 - 9.5303*10^-33 I`, which is the value of the complex number $e^{(2i\pi+1)^2}$. Completely different, wrong, and close to zero!

Here is how Equation (4.7) produces e as the value of $(e^{2i\pi+1})^{2i\pi+1}$. We start with

$$e^{i\pi} + 1 \stackrel{\text{Euler}}{=} 0, \Rightarrow (e^{i\pi})^2 = (-1)^2, \stackrel{*}{\Rightarrow} e^{2i\pi} = 1, \Rightarrow e^{2i\pi} e = e, \Rightarrow e^{2i\pi+1} = e. \tag{4.8}$$

(* Notice that we implicitly assumed that $(e^{i\pi})^2$ equals $e^{2i\pi}$. Fortunately it worked, but it is generally wrong.)

Based on this result, we denote $e^{(2i\pi+1)}$ by Z and $(2i\pi + 1)$ by W, and employ Equation (4.7) to produce

$$Z^W = (e^{2i\pi+1})^{(2i\pi+1)} \stackrel{(4.8)}{=} e^{(2i\pi+1)} \stackrel{(4.7)}{=} e^{W \ln Z} = e^{(2i\pi+1)\ln e} = e^{(2i\pi+1)} \stackrel{(4.8)}{=} e.$$

■ Similarly, exponentiation is not associative. The expressions a^{b^c} and $(a^b)^c$ are not the same. The first equals $a^{(b^c)}$, while the second is a^{bc}. The difference becomes crucial in the case of complex numbers, and here is a simple example $(2^2)^3 = 2^6$, but $2^{2^3} = 2^8$.

The case of complex numbers is, well, more complex, and here is an extreme example that illustrates this property clearly. We start with the well-known Euler equation

$$e^{i\theta} = \cos(\theta) + i \sin(\theta), \tag{4.6}$$

which implies the relations

$$i = e^{i\frac{\pi}{2}}, \quad i^i = \left[e^{i\frac{\pi}{2}}\right]^i = e^{-\frac{\pi}{2}}.$$

These relations are now used to compute two triple exponentiation. First comes

$$i^{i^i} = i^{(i^i)} = \left(e^{i\frac{\pi}{2}}\right)^{e^{-\frac{\pi}{2}}} = e^{i\frac{\pi}{2}e^{-\frac{\pi}{2}}} = \cos\left(\frac{\pi}{2}e^{-\frac{\pi}{2}}\right) + i\sin\left(\frac{\pi}{2}e^{-\frac{\pi}{2}}\right).$$

This is a messy expression that cannot be further simplified. On the other hand, the second triple exponentiation is simply i,

$$(i^i)^i = \left(e^{-\frac{\pi}{2}}\right)^i = e^{-\frac{\pi}{2}i} \overset{\text{argand}}{=} -i,$$

(the last term can be verified from the Argand diagram). An alternative derivation of this term is

$$(i^i)^i = i^{i^2} = i^{-1} = 1/i = -i.$$

■ The feature described here, originally discussed in reference [MindYourDecisions 17], is more a warning than a restriction. We start with the well-known rule of integration by parts

$$\int udv = uv - \int vdu,$$

And apply it to the functions $u = 1/x$ and $v = x$. The indefinite integral becomes

$$\begin{aligned}
\int \frac{1}{x}dx &= \left[\frac{1}{x}\right]x - \int x\left[\frac{-1}{x^2}\right]dx, \\
\int \frac{1}{x}dx &= \left[\frac{1}{x}\right]x - \int \frac{1}{x}dx, \\
\int \frac{1}{x}dx &= 1 + \int \frac{1}{x}dx.
\end{aligned}$$

This implies that $0 = 1$, an absurd result.

The reason for this error is our omission of the integration constant. A definite integral must be associated with a constant whose value may initially be unknown, but may later be determined from the context of the problem on hand. Once we realize that the indefinite integral of $f(x)$ is of the form $F(x) + C$, the expression above gets the form $F(x)+C_1 = 1+F(x)+C_2$ or $C_1 = 1+C_2$. All that we know about these constants is that their difference is one unit, but there is no error and no contradiction.

Notice that this applies only to indefinite integrals; a definite integral has no associated constant.

■ The problem of the repeating 9's is discussed and proved on Pages 6 and 100. It is also shown that proof #6 on Page 100 is wrong, because it swaps two limits. In general, swapping two consecutive limits, sums, derivatives, or integrals may lead to wrong results.

$$\lim_{m\to\infty}\left(\lim_{n\to\infty} 0.\underbrace{99\ldots9}_{m}{}^{n}\right), \quad \int_a^b\int_c^d dx\,dy, \quad \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}, \quad \frac{\partial}{\partial x}\frac{\partial}{\partial y}.$$

■ Finally, consider the simple, common question that many school children worry about, ponder, and ask, but rarely receive a full, straight answer. Why is the product of two negative numbers positive? Some answers are:

1. Because this convention is useful and doesn't lead to contradictions.

2. We can consider the negative sign a change of direction. Consider the product $(-3)(-5)$. The -3 reverses direction and leads us to the left side of the number line, and then the -5 reverses direction again and takes us back to the positive side. This explanation is faulty, since it doesn't explain why the sum of two negatives is negative.

3. Two negatives cancel each other out. This doesn't explain why they cancel?

Here is a simple explanation, based on elementary algebra

$$(-x)(-y) = (-x)(-y) + x \cdot 0 = (-x)(-y) + x(-y + y)$$
$$= (-x)(-y) + x(-y) + x \cdot y = -y(-x + x) + x \cdot y = x \cdot y.$$

Simple and convincing, while also not interfering with the rules of adding/subtracting negatives.

4.9 Debates in Mathematics

It is often surprising to hear that mathematicians, like theologians, philosophers, and politicians, have debates and doubts. These are caused by the fact that certain mathematical concepts can be used and understood in different ways. Here are a few examples.

■ The expression 0^0, which in Section 4.8 is claimed to be invalid, is nevertheless considered valid by certain mathematicians.

The reason is the smooth, continuous behavior of the function $f(x) = x^x$ in the vicinity of $x = 0$. As x approaches zero from the right, this function drops to a minimum at $x = 1/e$, and then starts climbing up, approaching the value 1 when x gets toward zero. This behavior is why some math enthusiasts claim that it is reasonable as well as useful to agree that $\lim_{x\to 0} f(x) = 1$, as long as it is also stated clearly that $f(x)$ features a discontinuity at zero.

■ Is $\sqrt{1}$ equal ± 1 or just 1? Those who believe in the former result claim that we always have to list ALL the roots of a number, but their opponents use the identity below (or a similar identity) and point out that claiming that $\sqrt{1} = \pm 1$ is equivalent to claiming that the $\sqrt[3]{8}$ is 2 and $(1 + i\sqrt{3})$ and $(1 - i\sqrt{3})$, which is ridiculous and makes things unnecessarily complicated.

The identity is

$$(-1 + i\sqrt{3})^3 = (-1 + i\sqrt{3})^2(-1 + i\sqrt{3}) = -2(1 + i\sqrt{3})(-1 + i\sqrt{3}) = 8,$$

and similarly for $(-1 - i\sqrt{3})^3$.

■ The problem of the repeating 9's. The repeating real number $0.999\ldots$—which is also denoted by $0.\bar{9}$, $0.\dot{9}$, and $0.(9)$—has long been recognized by mathematicians as equal to 1. Pages 6 and 100 list several proofs of this unintuitive fact. However, some people, mostly students of math, object to this result, often citing the following reasons:

1. I got used to thinking that a number must be represented uniquely. The claim that a number, let alone the familiar number 1, can have more than one representation seems to me a paradox.

2. In my mind, I see the notation 0.999 as a very long string of 9's. Its length is unspecified and may even be variable, but I can still envision that last 9 at infinity, and this prevents me from agreeing to the unexpected result $0.\bar{9} = 1$.

3. My intuition and experience tell me that $0.999\ldots$ is a sequence, and the limit of a sequence is an infinite process, not a well-defined value. After all, not every sequence has a limit.

4. The proofs that $0.999\ldots$ equals 1 are wrong. The number $0.999\ldots$ may not even exist, and even it if does, we cannot assume that $10 \cdot 0.999\ldots = 9.999\ldots$, because experience shows that we cannot simply extend arithmetic operations when infinities are involved. When we multiply $0.999\ldots$ by 10, the 9 to the left of the decimal point may not even be a full 9. It may be infinitesimally smaller than 9.

5. I can prove that $0.\bar{9} = 1$, but only in the system of surreal numbers.

6. Let's examine one of the traditional proofs, but this time use the digit 4 instead of 9. Consider the following:

$$x = 0.\bar{4}, \quad 10x = 4.444\ldots, \quad 9x = 10x - x = 4, \Rightarrow x = 4/9,$$

but direct multiplication shows that $0.44\ldots4 \times 9 = 3.99\ldots96$ regardless of how many 4's there are. Therefore, even if there are infinitely many 4's, there would still be a 6 at the "end."

■ A mathematical expression is executed by following the rules of precedence which are: (1) Execute anything in parentheses, (2) execute multiplications and divisions from left to right, and (3) execute any remaining additions and subtractions, also from left to right. A problem arises when a pair of parentheses include a number without any arithmetic operations, and the classic example is the expression $12 \div 3(4)$. Most people ignore the parentheses and end up with 16, but if we assume that the parentheses imply $3 \cdot 4$, then the result is 1.

> "This is incredible," I said, unable to locate the words to express my awe. It was indeed beyond incredible—an ineffable glimpse of the infinite.
>
> —Kane X. Faucher, *The Infinite Library.*

Infinity: A Poem

Infinity, a boundless sea,
A concept vast, eternally,
It stretches far beyond our sight,
A paradox both day and night.

In Zeno's race, Achilles chases,
The tortoise slow, yet space he paces,
For every step, a step remains,
In endless loops, Achilles strains.

A journey split in halves again,
Each smaller step, a quest in vain,
Infinity within a mile,
A paradox that makes us smile.

Cantor's sets, so large, profound,
In nested forms, infinity's found,
A countable infinity,
Yet larger still, uncountably.

In Hilbert's hotel, rooms galore,
Infinity opens up more doors,
With every guest that checks right in,
A paradox begins to spin.

The finite mind can't quite contain,
The endless stretch, the infinite chain,
Yet math and thought, they dare to stride,
In infinite realms, where truths reside.

The universe, its vast expanse,
In starlit skies, infinity's dance,
From galaxies to grains of sand,
Infinity holds all in hand.

In Mandelbrot's fractal lands,
Each pattern, endless, complex stands,
A simple rule, a boundless view,
Infinity in form anew.

Black holes with singularity,
Where time and space cease clarity,
An infinite dense point they say,
Where known laws bend and melt away.

Yet in our hearts and in our minds,
Infinity in dreams we find,
A symbol of the vast unknown,
A concept we may never own.

But still we strive to understand,
This endless stretch, this boundless land,
In paradoxes, truths are spun,
Infinity, the never done.

So ponder well this endless theme,
In finite thoughts, an infinite dream,
For in this quest, we touch the stars,
Infinity forever ours.

By `chatGPT 4o`, May 2024.

To see a world in a grain of sand
and a heaven in a wild flower,
hold infinity in the palm of your hand
and eternity in an hour.

—William Blake, *Auguries of Innocence,* 1810.

5
Order: Sequences and Series

Perhaps the first "equation" that we encounter, in arithmetic class in elementary school, is actually an identity

$$1 + 1 = 2.$$

This equation is taught in order to illustrate the symbols for plus, equal, and 2. It is, however, important because it also embodies the very essence of an equation and illustrates the format of equations. An equation has two parts, left and right, separated by the all-important equal sign. It tells us that when we compute the values of the expressions on both sides of this sign we will find that they are identical (or, if any quantities are unknown, we can assume that the two sides are equal). Once the child understands this simple equation, he is capable of understanding other, more complex, equations, not only purely mathematical equations, but equations from other fields of science and technology.

It seems that equations arise naturally when one tries to make sense of the many events, phenomena, and occurrences in the world around us. A case in point is Newton's all-important second law of motion $F = ma$. It appears as an equation as soon as we realize the relation between mass, acceleration, and force, and it practically begs to be written succinctly as a mathematical equation.

The English word "equation" is derived from the Latin *aequare*, meaning to make even or to level. Words such as equilibrium, adequate, egalitarian, equator, and equivalence are also derived from the same source.

The importance of equations in mathematics becomes clear when they incorporate unknown quantities. Even the most mathematically non-savvy person may sometimes ask himself, My chicken lays an egg a day. How many eggs a day would I have if I had five chickens?

◇ **Exercise 5.1:** (Funny.) If a chicken and a half lays an egg and a half in a day and a half, how many eggs does one chicken lay in one day?

5.1 Equations

A mathematical equation is an expression that contains—in addition to constants, variables, and operations—an equal sign and one or more unknowns. Not every equation is meaningful, and many equations seem to have no solutions. Given a set of equations, some may contradict others. Here are some examples.

The equation $x = x + 1$ is meaningless and has no solutions. If we subtract x from each side, we end up with $0 = 1$, a contradiction. The simple equation $x^2 = -1$ seems to have no solutions. No number that we know will satisfy it. Still, such equations occur often in mathematics and in science, which is why mathematicians decided to extend the concept of numbers and introduce the so-called complex numbers. The system of two equations in two unknowns $x = y + 1$ and $x = y - 1$ is contradictory. When we substitute the first in the second, the result is $y+1 = y-1$ or $+1 = -1$, a contradiction. The system of equations $x = y + 1$ and $2x = 2y + 2$ cannot be solved uniquely, because the two equations are equivalent; they say the same thing. Such a system has infinitely many solutions, because every value of x has a value of y that satisfies the equation.

◇ **Exercise 5.2:** Does the following equation have a solution or is it contradictory or even meaningless (unsolvable)?

$$\frac{x-1}{1-x} = 1.$$

A first-degree (linear) equation has the form $ax + b = 0$, where a and b are known numbers; they are the coefficients of the equation and x is the unknown. The solution of such an equation is trivial $x = -b/a$, which also implies that a must be nonzero.

A degree-2 (quadratic) equation has the form $ax^2 + bx + c = 0$. This type is also easy to solve and the solution $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$, has been known for centuries. Because of the $\pm$, such an equation has two solutions, but they may be identical (and may also be complex). Other ways to solve quadratic equations are described in references [WikiQuadr 24] and [YoutubePeyam 20].

A degree-3 (cubic) equation has the form $ax^3+bx^2+cx+d = 0$. It took many years of effort by several mathematicians to come up with the solution to these equations. Today, their fascinating efforts are the topic of several books and articles such as [Guilbeau 30] and chapter 6, pp. 133–154 of [Dunham 90].

◇ **Exercise 5.3:** The terms quadratic and cubic refer to a square and a cube, respectively. Why are these terms used to describe degree-2 and degree-3 equations?

A degree-4 (quartic, sometimes also called biquadratic) equation has ax^4 as its dominant term. In special cases where certain terms are missing, this equation can be solved by reducing it to a lower-degree equation. In the general case, a solution can be obtained by means of complex formulas that have been developed in the 1500s.

The problem starts when we reach degree 5 (quintic) equations. For several centuries mathematicians had looked in vain for a solution. Special cases of the quintic can be solved algebraically by applying a finite number of additions, subtractions, multiplications, divisions, and root extractions, the so-called *radicals* (from the name of the square root symbol $\sqrt{\ }$). However, no one was able to develop a similar solution for the general quintic equation. In the 19th century, Henrik Abel proved rigorously that equations of degree-5 and higher cannot be solved by means of radicals. Today, such equation can be solved numerically with the use of special functions that are infinite sums.

The celestial multiplication table.

A hundred thousand million stars
make one Galaxy;
A hundred thousand million galaxies
make one Universe.

—Arthur Eddington, *The Expanding Universe*, 1933.

5.2 The Pythagorean Theorem

On the face of it, the Pythagorean theorem states a simple property. Given a right triangle with sides a and b and a hypothenuse c, this theorem states that the square c^2 of the length of the hypothenuse equals the sum $a^2 + b^2$ of the square lengths of the two sides. Thus,

$$c^2 = a^2 + b^2.$$

Simple, but not trivial. When this theorem is taught to school children, most accept it without a murmur, but a few become puzzled by the fact that this property is satisfied by *every* right triangle. Those same children may also be puzzled by another basic fact, the relation $C = 2\pi R$ between the radius R of a circle and its circumference C. Why is it that *every* circle satisfies this relation, how can we be so sure about it, and why the number π? What is so special about this number?

The Pythagorean theorem is named after Pythagoras, because tradition has it that he was the first to prove it. Today it seems certain that this simple relation was known to the ancients long before Pythagoras's time. Ancient farmers, carpenters, and other professionals knew that a triangle with sides 3, 4, and 5 units creates a very precise right triangle. In modern times, the Pythagorean theorem found many applications in construction, manufacturing, and surveying.

The sum of the square roots of any two sides of an isosceles triangle is equal to the square root of the remaining side. Oh joy, rapture! I've got a brain!.

—Scarecrow in *The Wizard of Oz*, by Frank Baum.

Hundreds of proofs of the Pythagorean theorem have been developed over the centuries. Most are based on algebraic and geometric approaches, but some employ techniques from vector analysis, and even physics concepts such as mass and momentum.

Elisha Loomis started collecting such proofs in early 20th century and ended up with 370 of them, all fully described in [Loomis 68].

In addition to its many proofs, this celebrated theorem has been extended from right triangles to arbitrary triangles, where it is termed the law of cosines and its form is

$$c^2 = a^2 + b + 2 - 2ab\cos\theta,$$

where θ is the angle between sides a and b.

The chief question of interest to us is why is this particular theorem so popular and well known? It seems that the answer has several parts (1) it is one of the first theorems taught at school with a proof, (2) it has so many applications that we keep hearing of it all the time, and (3) it is easy to prove in many ways, based on different approaches.

Finally, a few proofs are listed, for the benefit of those who have the patience to follow them.

1. (Algebraic proof 65 from Loomis, page 65.) Given the right AHB triangle of Figure 5.1a, we copy it to form the rectangle AHBD. We drop a perpendicular from point H to locate a new point C. The two right triangles HDA and HBC are similar, so we conclude that HD/HB = AD/CB and HD × CB = HB × AD. Similarly, triangles DHB and AHC are similar, which leads to HD × AC = AH × DB. Adding yields $HD \times AB = HB \times AD + AH \times DB$.

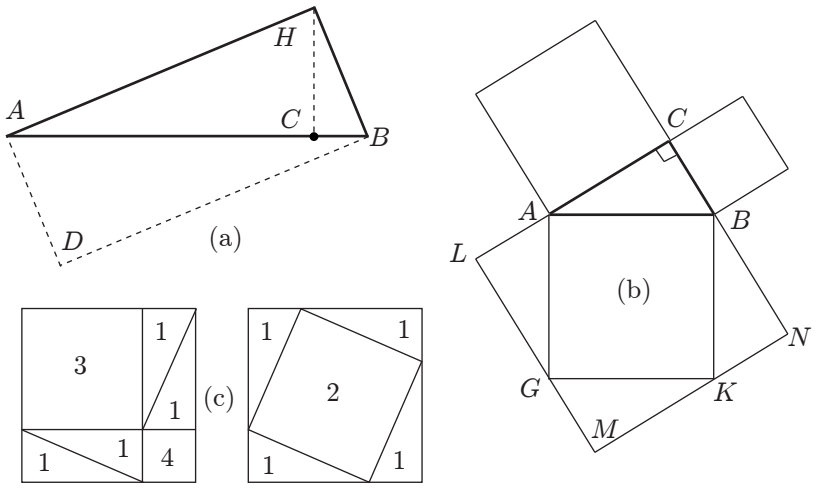


Figure 5.1: Proofs of the Pythagorean Theorem.

2. (Algebraic proof 69 from Loomis, page 67.) Figure 5.1a also implies that AB × BH = AH × DB + AD × BH or $AB^2 = HA^2 + HB^2$.

3. (Geometric proof 220 from Loomis, page 224.) Figure 5.1b shows a geometric construction based on the right triangle ABC. The hypothenuse AB is squared to determine points G and K. A line parallel to CB is drawn through G and another line, parallel to AC, is drawn through point K. These segments meet at the new point M. We now extend AC to meet segment MG at point L and extend segment CB to meet

line MK at N. This process forms the large square CLMN. It is now easy to see that the area of this square satisfies

$$\begin{aligned} CLMN &= 4\cdot\frac{1}{2}\cdot CB\cdot CA = (CB+CA)^2 - 2\cdot CB\cdot CA \\ &= CB^2 + 2\cdot CB\cdot CA + CA^2 - 2\cdot CB\cdot CA = AB^2. \end{aligned}$$

Therefore $AB^2 = BC^2 + CA^2$; easy!

4. This is a visual proof, already known to the ancient Chinese, to which some might object because it may not be rigorous. Figure 5.1c shows two identical large squares. The square on the left is partitioned into four identical right triangles labeled 1. This leaves the two squares labeled 3 and 4 on the sides of those triangles. The large square on the right is dissected into the same four right triangles (also labeled 1) and a smaller square, labeled 2, of the hypothenuses of these triangles. We now remove the eight triangles from the two large squares, which leaves the smaller squares 3 and 4 on the left and square 2 on the right, thereby showing that the combined areas of 3 and 4 equals the area of square 2.

Reference [Maor 07] is an extensive discussion of the Pythagorean theorem, its history, importance, and applications.

5.3 A Different Dirac Equation

Paul Dirac was an eminent theoretical physicist whose chief, but by no means only, claim to fame is the relativistic equation of the electron. The Dirac equation described here is different. It is much simpler, easy to understand, but is no less ingenious and unintuitive.

Around 1929, the following game was popular at Götingen University. Express any integer n using only basic arithmetic operators and exactly four occurrences of the digit 2. The first few integers are easy

$$1=\frac{2+2}{2+2},\quad 2=\frac{2}{2}+\frac{2}{2},\quad 3=2\times 2-\frac{2}{2},\quad 4=2+2+2-2,$$
$$5=2+2+\frac{2}{2},\quad 6=2\times 2\times 2-2,$$

but the next integer $7 = (2\times 2)!! - 2/2$, requires the double factorial operator !! that is defined by $n!! = n(n-2)(n-4)\ldots$. Obviously, one may claim that !! is not a basic operation, which is why even this (supposedly rigorous) mathematical game may suffer from arguments, disagreements, and misunderstandings.

When Dirac was introduced to this game, he thought about it intensely for a while, and then came up with the following general equation for any integer n

$$n = -\log_2\log_2\left[\underbrace{\sqrt{\sqrt{\ldots\sqrt{2}}}}_{n}\right].$$

Again, one could claim that this expression (the right-hand side of the equation above) employs only three 2's and an n, and that logarithms and/or square roots are not basic arithmetic operations, but the equation is nevertheless interesting and ingenious. (Another 2 can be included by writing $\sqrt[2]{2}$.) Because of human nature, this equation killed the game (see page 75 of [Casimir 83]). Here we will explain how it works.

The operation $\sqrt{2}$ can also be written $2^{1/2}$, which is why $\sqrt{\sqrt{2}}$ equals $2^{\frac{1}{2}\cdot\frac{1}{2}}$ or $2^{(\frac{1}{2})^2}$ and

$$\underbrace{\sqrt{\sqrt{\ldots\sqrt{2}}}}_{n}=2^{(\frac{1}{2})^n}.$$

In general, computing a logarithm requires many calculations, but the logarithm of an expression of the form 2^b is easy to figure out. One way to define logarithms is as follows. The logarithm of the number a is the number b that satisfies $10^b = a$. Similarly, the base-2 logarithm of a satisfies $2^b = a$. Therefore, the base-2 logarithm of $2^{(\frac{1}{2})^n}$ is $(\frac{1}{2})^n$ or 2^{-n}, and the Dirac equation becomes

$$-\log_2\log_2\left[\underbrace{\sqrt{\sqrt{\ldots\sqrt{2}}}}_{n}\right]=-\log_2(2^{-n})=n$$

> It is impossible to be a mathematician without being a poet in soul.
> —Sofia Kovalevskaya.

5.4 Sequences

A mathematical sequence is an ordered (finite or infinite) set of elements a_1, a_2, ..., together with a rule that determines the next element. The rule for the current element may be explicit (for example, $a_i = i^2$), it may depend on previous elements of the sequence (such as $a_i = a_{i-1} + a_{i-2}$), or may be anything else, even nonmathematical. Thus, 1, 4, 9, 16,... is the sequence of squared integers, and 1, 1, 2, 3, 5, 8, 13, ... is the well-known Fibonacci sequence. A series is the sum of the elements of a sequence. We start with two examples that illustrate the endless variety of these mathematical objects and the ingenuity of mathematicians working with sequences and series.

■ The infinite sequence 1, 3, 4, 6, 8, 9, 11, 12, 14, 16, 17, 19, 21, 22, 24, 25, 27, 29,... is an example of what mathematicians tend to consider beautiful. To find out the rule of this sequence, notice that its elements are separated by either 1 or 2, as the following figure shows

1 3 4 6 8 9 11 12 14 16
2 1 2 2 1 2 1 2 2.

Notice that the 2s on the bottom row of the figure are located in positions 1, 3, 4, 6, 8, 9,... of the row, but these positions are also the elements of our sequence. This simple

observation provides the key to this sequence. The next element a_{i+1} equals a_i+2 if i is an element of the sequence; otherwise $a_{i+1}=a_i+1$. This neat sequence is the brainchild of Daniel Mathews [Mathellaneous 08].

■ Given the infinite sequence 1/2, 2/4, 3/8, 4/16, 5/32, 6/64, 7/128, ... find its sum. This seemingly complex task becomes trivial when we write the sum in the special way shown here

$$\begin{aligned}
1/2+1/4+1/8+1/16+1/32+1/64+1/128+\cdots\\
1/4+1/8+1/16+1/32+1/64+1/128+\cdots\\
1/8+1/16+1/32+1/64+1/128+\cdots\\
1/16+1/32+1/64+1/128+\cdots\\
1/32+1/64+1/128+\cdots\\
1/64+1/128+\cdots\\
1/128+\cdots
\end{aligned}$$

The elements of the top row add up to 1. The next row is identical except that it is missing the term 1/2, so its sum is 1/2. The next row is missing 1/4, so its sum is 1/4, and so on. Thus, the sum of the original infinite sequence is $1+1/2+1/4+1/8+1/16+1/32+1/64+1/128+\cdots=2$.

Note. It is easy and elegant to show why the infinite sum $1/2+1/4+1/8+1/16+1/32+1/64+1/128+\cdots$ equals 1. We denote this sum by s and denote $a=1/4+1/8+1/16+1/32+1/64+1/128+\cdots$. It is now clear that $s-a=1/2$, and also that $s=2a$. These two equations are easily solved to yield $s=1$.

A juggler sequence is an integer sequence that starts with a positive integer a_0 and computes the next term by the recurrence relation

$$a_{k+1}=\begin{cases}\lfloor a_k^{1/2}\rfloor, & \text{if } a_k \text{ is even,}\\ \lfloor a_k^{3/2}\rfloor, & \text{if } a_k \text{ is odd.}\end{cases}$$

The name of this sequence reflects its behavior, which tends to go up and down as we move along the terms. Thus, if $a_0=3$, the first few terms are 5, 11, 36, 6, 2, and 1. Notice that 1 terminates a juggler sequence because $a_k=1$ implies that a_{k+1} is also 1.

See [juggler 09] for more information.

5.5 Numerical Sequences

A mathematical sequence is an ordered set of mathematical objects. If the set is finite, its elements may simply be stated. More interesting are infinite sequences, where the members of the sequence must be defined by an expression. Sequences, especially of integers, have been discovered, researched, and analyzed for a long time. In 1964, Neil Sloane started collecting such sequences and compiling the Encyclopedia of Integer Sequences [Sloane 08], which is currently located at `eois.org` and is searchable. This section describes a number of interesting and well-known sequences.

When faced with the first few elements of a sequence, the first questions that come to mind is What is the next element? What is the rule of the sequence? It is my belief

that mathematics, which is logical and consistent, also offers many pitfalls, so my first advice to anyone interested in integer sequences is, beware! Be careful. Don't make quick assumptions, and stay away from traps. The following quotes, from Richard Guy, should always be kept in mind "Superficial similarities spawn spurious statements," "Early exceptions eclipse eventual essentials," and "Capricious coincidences cause careless conjectures."

The Moser's circle problem and sequence, named after Leo Moser, is an example of such a trap. The following sequence should look familiar 1, 2, 4, 8, 16, ... Its elements are consecutive powers of 2, and it seems reasonable to guess that its next element is 32. However, there is another, similar but substantially different, sequence that starts in the same way, but continues with 31, 57, 99, 163, 256, 386,... (OEIS sequence `A000127`, reference [Sloane 08]). This is the little-known Moser's sequence, which results from dissecting the circle into regions by drawing chords. Figure 5.2 illustrates six examples. The original circle consists of a single region. Two distinct points are selected on the circle and a chord drawn between them which partitions the circle into two regions.

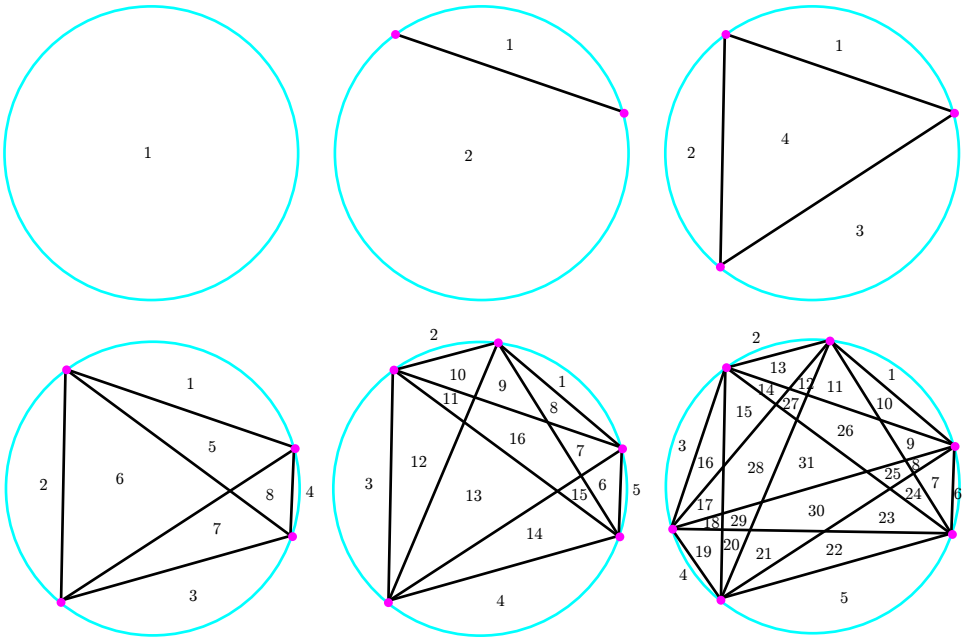


Figure 5.2: Circles and Rings.

When a third, distinct point is chosen and is connected to the other two by chords, the number of regions grows to four. The fourth point, with the help of more chords, results in eight regions. This grows to 16 regions when a fifth point is chosen, but point 6 increases the number of regions to 31 instead of the expected 32; a surprise! This surprise is an example of the strong law of small numbers, which is mentioned in the answer to Exercise 7.1.

Sometimes, when three chords happen to intersect at a point, the (normally small) triangular region that would have been defined by their intersections consists of only the single point of intersection and cannot be counted as a region. We therefore assume that the points are selected in such a way that no three chords intersect at a point.

Reference [youtube 23] analyzes the Moser circle sequence in detail, and uses the Pascal triangle to show that its elements are given by the simple relation

$$F_n = 1 + \binom{n}{2} + \binom{n}{4}.$$

The Golomb sequence. This clever sequence starts with 1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, ... Its general term a_n is the number of times n appears in the sequence. Thus, $a_1 = 1$ implies that 1 appears in the sequence once, while $a_2 = 2$ means that the integer 2 appears twice. Also $a_3 = 2$ tells us that 3 also appears twice. It is as if this sequence knows about itself. A little thinking shows that bigger integers appear more and more times in this sequence, and it has been shown that its general term a_n approaches $\phi^{2-\phi} n^{\phi-1}$ as n grows without limit.

The aliquot sequence. This is an intriguing sequence which can start with any positive integer. The successor of x in this sequence is the aliquot sum of x, the sum of its proper divisors (its divisors except the number itself). Thus, for example, the divisors of 24 are 1, 2, 3, 4, 6, 8, and 12, which is why their sum 36 is the aliquot successor of 24. Similarly, the successor of 36 is the sum of its divisors $1+2+3+4+6+9+12+18 = 55$, whose divisors are 1, 5, and 11. Thus, their sum 17, is the next term in the aliquot sequence 24, 36, 55, and 17, and is followed by 1, which terminates this sequence, because 17 is prime, its only divisor is 1, and the aliquot successor of 1 is zero. The main reference is either the wikipedia article on aliquot sequence or [aliquot 24].

We say that 24 is an abundant number because it is followed by a bigger number, while 55 is a deficient number because its aliquot successor is smaller. In general, the aliquot sequence of a number can end at 1, or in a loop. The latter case happens when the sequence reaches a perfect number or an amicable number. It may also happen that the sequence would never end, but such cases are currently unknown, because a long aliquot sequence tends to reach extremely large numbers from time to time, and computing the divisors of such numbers is time consuming and requires specialized software.

(Amicable numbers are two different natural numbers related in that the sum of the proper divisors of each equals the other. Examples are 220 and 284.)

The aliquot successor of a perfect number, such as 6, 28, or 496, is the number itself. The aliquot sequence of an amicable number A leads to its partner and from there back to A. There is also the case of sociable numbers. This is an aliquot sequence that forms a loop, such as 12496, 14288, 15472, 14536, and 14264. Some mathematicians propose the title "aspiring number" to an integer (such as 95) whose aliquot sequence terminates at a perfect number.

The Catalan-Dixon conjecture suggests that no aliquot sequence is infinite. Such a sequence either drops all the way to 1 or ends with a loop.

The aliquot sequence of 138 oscillates many times before it reaches a maximum of approximately 1.79×10^{11} and then drops down to terminate after almost 180 steps.

Currently, in 2024, the integer 276 is the first whose aliquot sequence reaches numbers so large that it is impractical to continue and compute their entire sequence. Recall that the sequence depends on factoring integers, and there currently isn't a fast algorithm to do that. As a result, 276 is currently the first Lehmer number. Other Lehmer numbers, named after Derrick H. Lehmer, are 552, 564, 660, and 966. .

5.5.1 The 2, 3, 4, 82,000 Sequence

Having 10 fingers, we normally use decimal numbers, but Exercise 5.6 shows that any real number can serve as the base of a number system. We also know that base b requires b digits. Armed with these facts, we start with binary (base 2) numbers. They use only the two digits 0 and 1, which is why every binary number consists of only those digits. Ternary (base 3) numbers use the digits 0, 1, and 2, with is why 0 and 1 appear in such numbers less than 100% of the time. Base 4 numbers use 0, 1, 2, and 3, which is why the percentage of 0's and 1's in such numbers is even smaller.

Upon realizing this, mathematicians, being curious, may ask what integers are represented with only 0's and 1's in every base. Such integers are currently unknown, and the best that is known about this problem is summarized in Table 5.3, from reference [Grime 15].

Base	2	3	4	The integer 82,000
2	10	11	100	10100000001010000
3	2	10	11	11011111001
4	2	3	10	110001100
5	2	3	4	10111000

Table 5.3: The 2, 3, 4, 82,000 Sequence.

The smallest integers that are represented with only 0's and 1's in bases 2, 3, and 4 are 2, 3, and 4, respectively, but the smallest integers that are represented with only 0's and 1's in base 5 is 82,000; a surprise! Unfortunately, the same 82,000, when represented in base 6, proves to be $6^6 + 6^5 + 6^4 + 6^3 + 6^2 + 6^1 + 6^0 = 1431344_6$, a disappointment.

In spite of searches with fast computers, it is currently unknown whether there are integers that use only 0's and 1's in bases 2 through 6. The unusual sequence 2, 3, 4, 82000, known as sequence A258107 in the `OEIS`, therefore stands alone.

◇ **Exercise 5.4:** Why research and explore sequences like this, which seem to have no use, no purpose, no reason, and no end?

5.6 The Fibonacci Sequence

The Fibonacci sequence of integers F_n is based on the following simple recursive relation which is also general and is used by other sequences of numbers

$$G_n = G_{n-1} + G_{n-2} \tag{5.1}.$$

Technically, this is a rule of linear recurrence with constant coefficients. Since each term of the sequence depends on the two preceding terms, the first two terms of any sequence based on Equation (5.1) must be given explicitly. For the Fibonacci Sequence, these terms are either 0, 1 or 1, 1. Either choice generates the same sequence and the difference between them is the indexing of the terms. It is common to denote the first two items by $F_0 = 0$ and $F_1 = 1$. The sequence then becomes

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots.$$

Elements of this sequence that are also primes are referred to as Fibonacci primes. It is not known whether there are infinitely many Fibonacci primes, and this situation is common in number theory, where simple problems whose solutions have eluded even the greatest mathematicians are common.

A detailed mathematical reference is [WikiFibon 23], but there are many books and videos on this sequence and its relation to the golden ratio ϕ.

The Fibonacci Sequence was already known to Indian mathematicians around the third century B.C. It first appeared in Europe in the 1202 book *Liber Abaci* by Leonardo Fibonacci, who used it to describe the rapid growth of rabbit population. To a modern mathematician, the importance of this sequence is in its relation to the golden ratio ϕ (also denoted by φ).

It can be shown (see, e.g., reference [WikiFibon 23]) that the Fibonacci numbers can be represented explicitly by the expression

$$F_n = \frac{1}{\sqrt{5}}\left[\frac{1+\sqrt{5}}{2}\right]^n - \frac{1}{\sqrt{5}}\left[\frac{1-\sqrt{5}}{2}\right],$$

and this expression can be used to prove certain properties of the Fibonacci sequence, such as the following. The sum of the squares of the first k Fibonacci numbers equals the product $F_k \cdot F_{k+1}$, as is demonstrated by the Mathematica code:

```
k=8;
fn[n_]:=(((1.+Sqrt[5])/2)^n)/Sqrt[5] - (((1-Sqrt[5])/2)^n)/Sqrt[5];
Sum[fn[i]^2,{i,k}]
fn[k]fn[k+1]
```

The relation between φ and the Fibonacci numbers is easy to verify. One way to start exploring this relation is Binet formula, which provides the closed-form expression

$$F_n = \frac{(1+\sqrt{4})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}} = \frac{\phi^n - \psi^n}{\phi - \psi} = \frac{\phi^n - \psi^n}{\sqrt{5}} = \frac{\phi^n - (-\phi)^{-n}}{\sqrt{5}} = \frac{\phi^n - (-\phi)^{-n}}{2\phi - 1},$$

where ϕ and ψ are the golden ratio $(1+\sqrt{5})/2 \approx 1.61803\ldots$ and its conjugate $(1-\sqrt{5})/2 = 1-\phi = -1/\phi \approx -0.61803\ldots$, respectively.

Another way to illustrate this relation is to check, with the use of mathematical software, that the ratio F_{n+1}/F_n converges to ϕ as n approaches infinity. This has already been observed by Johannes Kepler. While not a formal proof, such a check is very convincing. The following is simple Mathematica code.

```
f1 = 1.; f2 = 2.;
Do[{Print[f2/f1], f = f1, f1 = f2, f2 = f2 + f}, {i, 15}]
```

◇ **Exercise 5.5:** Applicants to the math program at Hale university must first pass an Entrance Exam. When Prof. Tenib was asked to contribute a question to this exam, he recalled the term $(1-\sqrt{5})^n$ from the famous formula of his cousin Binet, and he came up with the following unusual question: Simplify the value of $(1-\sqrt{5})^8$ as far as possible without a calculator. As a potential future applicant to the same university, try your hand at this problem.

The original Fibonacci sequence starts with either 0, 1 or 1, 1, but it may come as a surprise to realize that the precise first two elements do not matter. Any two integers may serve as the first two elements of a Fibonacci-type sequence, satisfying Equation (5.1), provided that both are not zero.

Even more, given any sequence defined by the recursive relation $x_n = x_{n-1} + x_{n-2}$, if its elements satisfy $\frac{x_n}{x_{n-1}} = \frac{x_{n-1}}{x_{n-2}} = C$, then the constant ratio C is always φ. The proof is simple

$$x \stackrel{\text{def}}{=} \frac{x_n}{x_{n-1}} = \frac{x_{n-1}}{x_{n-2}} \Rightarrow \frac{x_{n-1}+x_{n-2}}{x_{n-1}} = \frac{x_{n-1}}{x_{n-2}}$$
$$\Rightarrow 1 + \frac{x_{n-2}}{x_{n-1}} = \frac{x_{n-1}}{x_{n-2}} \Rightarrow 1 + \frac{1}{x} = x \Rightarrow x^2 - x - 1 = 0,$$

which is the familiar equation whose solution is φ, a constant. This shows that the original Fibonacci sequence isn't special. What is special is the recursive rule on which it is based.

There are also M-Fibonacci (Mbonacci) sequences which are defined by the similar recursive relation $F_n = M \cdot F_{n-1} + F_{n-2}$.

The anti-Fibonacci numbers are defined recursively as $f(1) = 1$, $f(2) = 0$, and $f(k+2) = f(k) - f(k+1)$. The sequence starts with 1, 0, 1, −1, 2, −3, 5, −8, ... This sequence is also obtained when the Fibonacci sequence is extended backward from 0. Thus,

$$\ldots -8, 5, -3, 2, -1, 1, 0, 1, 1, 2, 3, 5, 8, \ldots.$$

In much the same way that the ratios of successive Fibonacci numbers converge to ϕ, the ratios of successive anti-Fibonacci numbers converge to $1/\phi$.

Reference [youtube.gen.fibo 24] discusses many ways of generalizing this important sequence, some of which are listed below.

- Fibonacci integer sequences. These are all the integer sequences that have the form $g(n+2) = g(n) + g(n+1)$ for any integers $g(0)$ and $g(1)$ that are not both zero. Such a

sequence can be expressed in terms of its first two elements as $g(n) = F_n g(1)+F_{n-1}g(0)$, where F_n are the Fibonacci numbers.

- The 3-Fibonacci sequence is $a(n) = 3a(n-1) + a(n-2)$, with $a(0) = 0$ and $a(1) = 1$. This is sequence A006190 in the OEIS. There are similar 4-, 5-, and 6-Fibonacci sequences.

- The (1,2)-Fibonacci sequence, which also known as the Jacobsthal sequence, is $a(n) = a(n-1) + 2a(n-2)$, with $a(0) = 0$ and $a(1) = 1$. This is sequence A001045 in the OEIS.

- The Tribonacci numbers are defined as $a(n) = a(n-1)+a(n-2)+a(n-3)$ for $n \geq 3$ with $a(0) = a(1) = 0$ and $a(2) = 1$. Mathematicians are meticulous and love to extend and generalize concepts and terms, which is why there are also tetranacci, pentanacci, hexanacci, heptanacci, octanacci, and enneanacci numbers (if you can believe that).

- The Fibonacci word is the letter-string analog of the Fibonacci numerical sequence. It is defined by:

$$F_n = \begin{cases} b & n = 0, \\ a & n = 1, \\ F_{n-1}F_{n-2} & n > 1. \end{cases}$$

Its first few elements are `b`, `a`, `ab`, `aba`, `abaab`, `abaababa`, `abaababaabaab`,... (sequence A106750 in the OEIS). Notice that the length of each element is the corresponding Fibonacci number.

- Keith number. This intriguing concept was introduced into the mathematical world by by Mike Keith in 1987. Start with a natural number n with k digits in some number base. Consider those digits the first k elements of a Fibonacci-like sequence in which each element is the sum of its k predecessors. If any element of this new sequence equals n, then n is a Keith number.

  Strange, rare, and difficult to find, Keith numbers were named by their "creator" repfigit numbers (short for repetitive Fibonacci-like digit). Some examples of decimal Keith numbers are 14, 19, 28, 47, 61, 75, 197, 742, 1104, 1537, 2208, 2580, and 3684. More are known in other number bases.

The Pisano period

One of the names by which Leonardo Fibonacci is known today is Leonardo Pisano. The period described here can be found, for example, in reference [youtube.pisano 13], and it has to do with the Fibonacci numbers. When successive Fibonacci numbers are divided by an integer, the remainders form a periodic sequence. The period depends on the dividend, and it is referred to as the Pisano period. Thus, the Mathematica statement `Table[Last[QuotientRemainder[Fibonacci[n],7]],{n,1,35}]` produces the period-16 sequence

```
{1,1,2,3,5,1,6,0,6,6,5,4,2,6,1,0,
 1,1,2,3,5,1,6,0,6,6,5,4,2,6,1,0,
 1,1,2}
```

Dividing the same Fibonacci numbers by 5 yields a sequence of remainders with a Pisano period of 20. The Pisano period is just one of the many curious properties of the Fibonacci sequence.

> Each wife of Fibonacci,
> Eating nothing that wasn't starchy
> Weighed as much as the two before her.
> His fifth was some signora!
> —James A. Lindon, In Martin Gardner, *Mathematical Circus*, 1979.

Applications of the Fibonacci sequence

Just as the golden ratio has fired the imagination of countless faithful who "see" in it all kinds of properties and features, so has the Fibonacci sequence. Out of all the strange beliefs and claims related to this sequence, I would like to mention the book [Mandalian 24], which claims that the Fibonacci numbers shape the duration of historical progress.

Out of the myriad of applications of the Fibonacci sequence only three will be mentioned here (1) converting between miles and kilometers, (2) the variable-length Fibonacci codes, and (3) the famous spirals found in sunflowers and their relation to the Fibonacci sequence.

1. Fibonacci Number base. We normally use decimal numbers, but only because we have 10 fingers. Mathematically mature people know that integers can be represented in other number bases, and 2, 8, and 16 are number bases commonly used by computer professionals. However, the base of a numbering system does not have to be an integer and can even be irrational. Fibonacci aficionados may be happy to know that ϕ can serve as the base of a numbering system of a numbering system, and numbers can easily be converted between this base and any other base. This useful fact is expressed by Zeckendorf's theorem which states Every positive integer can be expressed uniquely as a sum of distinct non-consecutive Fibonacci numbers, using just the two binary digits 0 and 1.

Converting an integer to Fibonacci base is straightforward. We assume that the Fibonacci sequence starts with $F_1 = 1$, $F_2 = 2$, $F_3 = 3$, $F_4 = 5$, etc. The Fibonacci base of an integer is a binary string where the weight assigned to each bit from right to left is the next Fibonacci number instead of the next power of 2. Thus, the Fibonacci code of $9 = 8 + 1$ is 10001, the code of $10 = 8 + 2$ is 10010, and that of $20 = 13 + 5 + 2$ is 101010. Also, the Fibonacci-base number 1000101 is 21+3+1=25.

An elegant application of Fibonacci-based integers is conversion between miles and kilometers. This is fortuitous, because one mile equal 1609 meters or 1.609 kilometers, a number very close to ϕ. Thus, if a number of miles is converted to the Fibonacci based 100101, it can be converted to kilometers simply by shifting it one position to the left 1001010. Naturally, shifting to the right converts in the opposite direction.

◇ **Exercise 5.6:** Can any real number serve as the base of a number system?

◇ **Exercise 5.7:** How can $0.666\ldots$ equal 1?

2. Fibonacci data compression. In our modern world, we depend on our computers and other digital equipment all of which deal with data, lots and lots of data. One way to deal with more and more of the data accumulating all the time is to compress data files that are not in use and keep them outside our computer, in a digital archive, or in a "cloud."

Data can be compressed because it is not random and its normal representation contains redundancies. An example is a text file. In many languages, certain letters tend to appear more often than others. In technical terms, the letter distribution is nonuniform. In English, for example, the letter `E` appears about 10–11% of the time, followed by `A` (8.5%), `I` (7.5%), `O` (7.1%), `T` (7%), and so on, down to `Z`, `J`, and `Q`, which are rare. Thus, a simple approach to text compression is to replace each letter with a variable-length code and assign short codes to the frequent letters and long codes to the rare letters.

It is easy to encode text in this way, but the codes must be designed to make them uniquely decodable. Here is an example. A four-letter alphabet is given, consisting of `A` (50%), `B` (25%), `C` (15%), and `D` (10%). After considering their frequencies we assign them the binary codes 1, 01, 11, and 010, respectively. Given the text `ABCD...`, we encode it as the bitstring 1|01|11|010... (without the vertical bars). When this is decoded, the first two codes are identified uniquely, but the following 11|010... is ambiguous, because it can also be decoded as 1|1|01|0....

Many methods and approaches have been developed for designing uniquely decodable variable-length codes, and one of them, reference [Fraenkel 96], employs Fibonacci numbers. A paragraph on Page 288 explains how the Fibonacci sequence can serve as the base of a numbering system where numbers are represented using just 0 and 1. Such a system has the following properties (1) the leftmost bit is always 1, and (2) there are no consecutive 1's. (The latter is because a code such as $10110 = 8 + 3 + 2$ must be replaced with $11000 = 8 + 5$, which is turn must be replaced with $100000 = 13$.)

To construct the variable-length Fibonacci code of the positive integer n, the binary Fibonacci-based representation of n is reversed, and an extra 1 is appended to the right end. Thus, the variable-length Fibonacci code of 5 becomes 0001|1 and that of 33 is 1010101|1. It is obvious that such a code always ends with a pair 11, and that this is the only such pair in the codeword, because the Fibonacci representation does not have adjacent 1's. This property makes it possible to decode such a code unambiguously, but also causes these codes to be long, because not having adjacent 1's limits the number of possible binary patterns.

Table 5.4 lists the first 12 Fibonacci codes.

1	11	7	01011
2	011	8	000011
3	0011	9	100011
4	1011	10	010011
5	00011	11	001011
6	10011	12	101011

Table 5.4: Twelve Fibonacci Codes.

To decode such a codeword, skip bits of the code until a pair 11 is reached. Replace this 11 by 1. Multiply the skipped bits by the values ..., 13, 8, 5, 3, 2, 1 (the Fibonacci numbers), and add the products. Obviously, there is no need to actually multiply; simply use the 1 bits to select the proper Fibonacci numbers and add them .

The Fibonacci codes are long, but have the advantage of being more robust than most other variable-length codes. A corrupt bit in a Fibonacci code may change a pair of consecutive bits from 01 or 10 to 11 or from 11 to 10 or 01. In the former case, a code may be read as two codes, while in the latter case two codes will be decoded as a single code. In either case, the slippage will be confined to just one or two codewords and will not propagate further.

The length of the Fibonacci code for n is less than or equal to $1 + \lfloor \log_\phi \sqrt{5}n \rfloor$.

In reference [Fraenkel 96], the authors refer to this code as C^1 and offer two variations named C^2 and C^3. Code C^2 is constructed from C^1 as follows:

1. Each codeword in C^1 ends with two consecutive 1's; delete one of them.
2. Delete all the codewords that start with 0.

Thus, the first few C^2 codewords, constructed with the help of Table 5.4, are 1, 101, 1001, 10001, 10101, 100001, and 101001. An equivalent procedure to construct this code is the following:

1. Delete the rightmost 1 of every codeword in C^1.
2. Prepend 10 to every codeword.
3. Include 1 as the first codeword.

A simple check verifies the equivalence of the two constructions. Code C^2 is not a straightforward Fibonacci code as is C^1, but it can be termed a Fibonacci code, because the interior bits of each codeword correspond to Fibonacci numbers. The code consists of one codeword of length 1, no codewords of length 2, and F_{r-2} codewords of length r for any $r \geq 3$. Individual C^2 codewords are identified by the decoder because each starts and ends with a 1. Thus, two consecutive 1's indicate the boundary between codewords. The first codeword introduces a slight complication, but can be handled by the decoder. A string of the form $\ldots 011110 \ldots$ is interpreted by the decoder as $\ldots 01|1|1|10 \ldots$, i.e., two consecutive occurrences of the codeword 1.

Code C^2 is also more robust than C^1. A single error cannot propagate far because the decoder is looking for the pattern 11. The worst case is a string of the form $\ldots xyz \ldots = \ldots 1|10 \ldots 01|1|10 \ldots 01|1 \ldots$ where the middle 1 gets corrupted to a 0. This results in $\ldots 1|10 \ldots 01\mathbf{0}10 \ldots 01|1 \ldots$ which is interpreted by the decoder as one long codeword. The three original codewords xyz are lost, but the error does not propagate any further. Other single errors (corrupted or lost bits) result in the loss of only two codewords.

Code C^3 is constructed from C^1 as follows:

1. Delete the rightmost 1 of every codeword of C^1.

2. For every $r \geq 1$, select the set of C^1 codewords of length r, duplicate the set, and distinguish between the two copies by prepending a 10 to all the codewords in one copy and a 11 to all the codewords in the other copy.

This results in the codewords 101, 111, 1001, 1101, 10001, 10101, 11001, 11101, 100001, 101001, 100101, 110001, ... It is easy to see that every codeword of C^3 starts with a 1, has at most three consecutive 1's (and then they appear in the prefix), and every codeword except the second ends with 01. The reference above proves that for any $r \geq 3$ there are $2F_{r-2}$ codewords.

The decoder first checks for the pattern 111 and interprets it depending on the bit that follows. If that bit is 0, then this is a codeword that starts with 111; otherwise, this

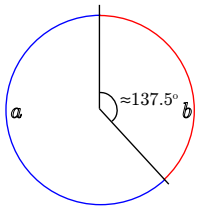
is the codeword 111 followed by another codeword. If the current pattern is not 111, the decoder checks for 011. Every codeword except 111 ends with 01, so the pattern 01|1 indicates the end of a codeword and the start of the next one. This pattern does not appear inside any codeword.

Given an r-bit codeword $y_1 y_2 \ldots y_r$ (where $y_r = 1$), its index (i.e., the integer whose codeword it is) is given by

$$\begin{aligned} &2F_{r-1} - 2 + y_2 F_{r-2} + \sum_{i=3}^{r} y_i F_{i-1} - F_{r-1} + 1 \\ &= \sum_{i=3}^{r+2} y_i F_{i-1} + (y_2 - 1)F_{r-2} - 1, \end{aligned}$$

where $y_{r+1} = 1$ is the leftmost bit of the next codeword.

3. Phyllotaxis, spirals, and sunflowers



The golden angle, the close cousin of the golden ratio, is constructed by partitioning a circle into two arcs a and b, such that $a > b$ and $a/b = \phi$. This happens when the angle between the radii defining the arcs is approximately 137.5°, as illustrated in the figure. The golden angle is often found in plants where it is the angle between successive leaves. Natural selection favors plants that receive more sunlight, which is why the arrangement of leaves around a plant is important. The worst phyllotaxis occurs when leaves grow under or over one another, thereby blocking the sun.

The value of the golden angle is easy to compute (see Page 64). The circumference of a circle is always 2π and in our case it is also $a + b$. The ratio of the arc subtended by b to the entire circumference is therefore

$$\frac{b}{2\pi} = \frac{b}{a+b} = \left[\frac{a+b}{b}\right]^{-1} = \left[1 + \frac{b}{a}\right]^{-1} = \frac{1}{1+\phi} = \frac{1}{\phi^2}.$$

Thus, exactly ϕ^2 golden angles fit in the 2π circumference of the circle, which results in

$$\text{Golden angle} = \frac{2\pi}{\phi^2} \approx 2.39996 \text{ radian, or } \frac{360}{\phi^2} \approx 137.508°.$$

Phyllotaxis is the study of the arrangement of leaves, petals, seeds, etc., on plants. It sometimes happens that as a plant grows, new leaves appear along its axes, facing in different directions. When a smooth imaginary curve is drawn, connecting the tips of successive leaves, the result is a spiral.

> Pick up a pinecone and count the spiral rows of scales. You may find eight spirals winding up to the left and 13 spirals winding up to the right, or 13 left and 21 right spirals, or other pairs of numbers. The striking fact is that these pairs of numbers are adjacent numbers in the famous Fibonacci series: 1, 1, 2, 3, 5, 8, 13, 21.... Here, each term is the sum of the previous two terms. The phenomenon is well known and called phyllotaxis.
>
> —Stuart A. Kauffman, *At Home in the Universe*, 1996.

Why do sunflowers have so many florets and why are the florets arranged in those beautiful spirals that seem to radiate and curve both clockwise and counterclockwise? We believe that this is because such packing of the florets gives today's sunflowers an advantage that past sunflowers did not have. Out of many species of competing sunflowers, natural selection, working blindly, has chosen the ones that managed to pack the most florets and thus to spread the most number of seeds around them.

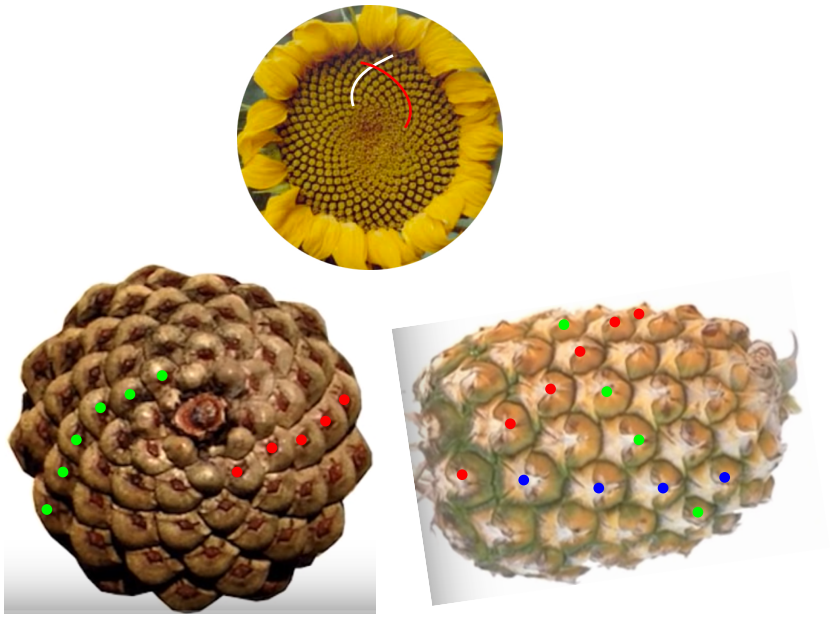


Figure 5.5: Pinecone Spirals.

Over time, botanists have counted both the number of sunflowers spirals and their lengths. This is not a simple task, since a sunflower is a living being, not an inanimate object. It is not perfect and it tends to make mistakes while growing. The botanists were surprised to discover that both the number of spirals and their lengths tended to be Fibonacci numbers.

Today, there is a simple mathematical model that explains the presence of Fibonacci numbers in sunflowers, as well as in other plants such as pinecones, aloe plants, the flowering of artichoke, and flower petals. The model assumes that as the sunflower grows and develops, new florets appear at its center and slowly move away radially as

new florets are created. The model further assumes that each new floret rotates about its center through a constant angle $2\pi\alpha$ before it takes each step on its way toward the edge of the sunflower. At a certain point, the sunflower reaches its adult size and its growth stops.

◇ **Exercise 5.8:** Search the Internet for examples of Fibonacci numbers in plants.

Slight random perturbations in a floret's speed and rotation angle α—due perhaps to variations in temperature, lighting, and moisture—may cause a new floret to move in a direction or at a speed slightly different from what nature intended, and thereby cause an imperfect final pattern. Even before this model is implemented and tested, it is easy to imagine that the value of α is critical and even tiny changes in it may result in substantially different paths for the florets and a hugely different final pattern at the end of the growth process.

This simple model is based on the original work described in reference [Vogel 79]. Once we agree on this model, it is natural to explore it by computer simulation, and vary its parameters (speed and rotation angle) in order to determine the values that produce floret patterns similar to those observed in nature. Rational values of α result in a pattern of straight rays radiating from the center, which is not what is actually observed in real sunflowers. Choosing $\alpha = n/m$, where n and m are positive integers with no common factors, produces a pattern of m straight rays, with empty spaces in between. The dots in Figure 5.6a were generated by $\alpha = 22/7$, and their serial numbers indicate how consecutive dots belong to adjacent rays. Dots 1, 7, 14, and 21 make up one ray while their successors, dots 2, 8, 15, and 22, constitute its neighbor ray. This is similar to the way the stars move from arm to arm in a spiral galaxy.

The next test should therefore try irrational values of α. Figure 5.6b is the pattern obtained for $\alpha = \pi$. It is surprisingly similar to that of $\alpha = 22/7$, but here we finally see spirals instead of straight rays, and this is because π is slightly less than $22/7$. To understand this behavior, examine dot 21 in the straight ray formed by dots 1, 7, 14, and 21 in Figure 5.6a. Its successor, dot 22, is located on the next straight ray, but if the rotation angle α of dot 21 were slightly smaller (in particular, π instead of $22/7$), it would cover a shorter distance on its way to the next ray and would cause that ray to bend. When this argument is applied to all the points, the bending of an arm becomes a spiral.

This is encouraging, but there still are spaces between adjacent spirals, which are caused by π being "not very irrational" (or not a "good" irrational). We know that there are infinitely many rational numbers that get closer and closer to any irrational number, and yet the discussion on Page 90 shows why the golden ratio φ can be considered the most irrational number.

We continue with the pattern of Figure 5.6c, which is produced by a rotation angle $\alpha = e$, the Euler constant. The spirals here are more pronounced, but the important feature is the extra spiral that suddenly appears between points 21 and 22. The final dot pattern in this case consists of two interleaved groups of four spirals each. There is less space between adjacent spirals, and we must conclude that this is because e is more irrational than π.

◇ **Exercise 5.9:** Prove that if the rational $c = a/b$ approximates the irrational $\sqrt{x}$, then there are infinitely many rationals closer to $\sqrt{x}$ than c.

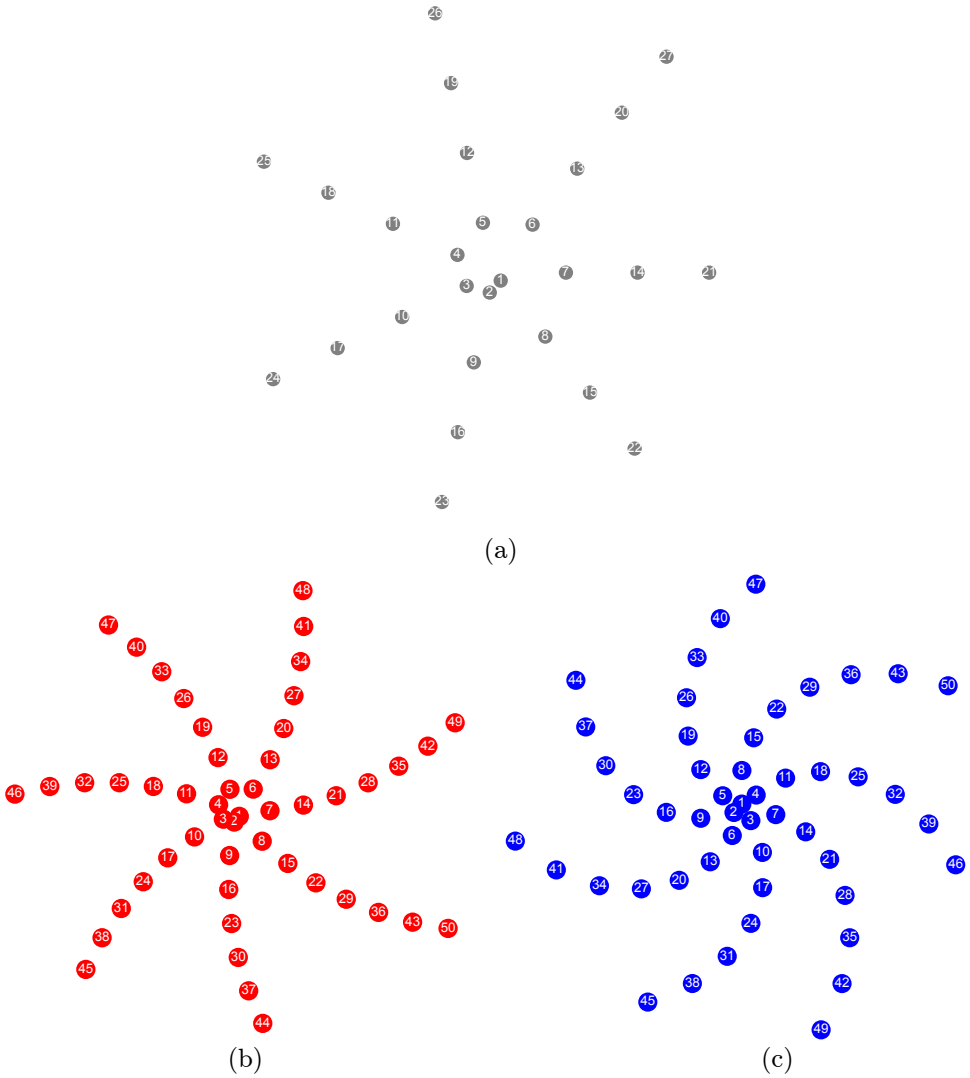
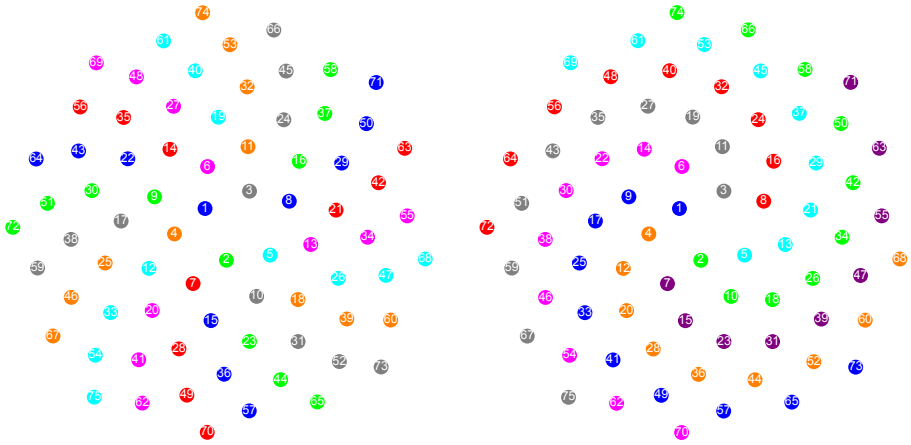


Figure 5.6: Simulated Sunflower Spiral Patterns.

Figure 5.7 employs the irrational rotation angle φ, and the resulting pattern justifies the claim that φ is the most irrational number. The two versions of the figure feature identical dot configurations, but the different colors make it obvious that the spirals can be chosen in different ways, and their number is always a Fibonacci number.



Seven colors Eight colors

Figure 5.7: Simulated Sunflower Spiral Patterns.

This observation supports Vogel's simple model. When the rotation angle is ϕ, the number of spirals in the simulation is always a Fibonacci number, because ϕ is the limit of the ratio F_n/F_{n-1} of Fibonacci numbers. Since ϕ is irrational, there are infinitely rational approximations that get closer and closer to ϕ, and any real world simulation or computation that involves ϕ must use one of them. Thus, our simulation uses the rational number F_n/F_{n-1} (for some unknown n) instead of the precise value of ϕ. This produces, as shown five paragraphs earlier, a final pattern with F_{n-1} straight rays, with empty spaces in between.

Figure 5.8 lists the Mathematica Code for Figures 5.6 and 5.7 (see also Figure 1.12). The rotation angle is computed on line 6, and the reader can experiment with this code and change `(Sqrt[5] + 1)/2` on this line to any other value.

Finally, Figure 5.9 shows spirals with lengths the are Fibonacci numbers.

The Supergolden ratio

The supergolden ratio, denoted by ψ, is defined, similar to the golden ratio, by the cubic equation $x^3 = x^2 + 1$. Its precise value is the real solution of this equation

$$\psi = \frac{1}{3}\left[1 + \sqrt[3]{\frac{29+3\sqrt{93}}{2}} + \sqrt[3]{\frac{29-3\sqrt{93}}{2}}\right] \approx 1.46557123187676802665673 12\ldots.$$

As with the association of ϕ with the Fibonacci sequence, ψ has its own supergolden sequence, also known as the Narayana's cows sequence. Its main feature is the ratio

```
(*Dots forming a sunflower  pattern*)
(*alpha= 2Pi*(the irrational phi)*)
lc = {};
clr = {Red, Blue, Green, Gray, Orange, Cyan, Magenta, Purple, Black};
Do[r = Sqrt[k];
 t = 2 Pi k (Sqrt[5] + 1)/2;
 kk = Mod[k, 8] + 1;
 AppendTo[
  lc, {clr[[kk]], Disk[{r Cos[t], -r Sin[t]}, 0.3], White,
   Text[k, {r Cos[t], -r Sin[t]}]}], {k, 75}]
Graphics[{Gray, lc}]
```

Figure 5.8: Mathematica Code for Figures 5.6 and 5.7.

between consecutive terms, which approaches ψ. The Narayana sequence is defined by $a_{n+1} = a_n + a_{n-2}$ with 1's as its first three elements. The first values are 1, 1, 1, 2, 3, 4, 6, 9, 13, 19, 28, 41, 60, 88, 129, 189, 277, 406, 595.... There is also a supergolden rectangle, whose sides have the ratio $\psi : 1$.

See also Page 161 for an extended Fibonacci sequence.

5.6.1 The Lucas Sequence

The Lucas sequence L_n, named after François Lucas is defined by the same recursive rule as Fibonacci, but with the first two elements being 2 and 1. Thus, it starts with 2, 1, 3, 4, 7, 11, 18, 29, 47, ... Elements of this sequence that are also primes are referred to as Lucas primes. For those who would like the full story, the closed formula for these elements is $L_n = \varphi^n + (-\varphi)^{-n}$.

The Lucas sequence is also often referred to as Lucas numbers, because it is easy to generalize it in many ways and end up with many different Lucas sequences.

Digression (a joke). We can imagine how Lucas came to this definition. "Now let's see," he says. "Fibonacci starts with 1, 1, 2, 3, so why not start my new sequence with 1, 2? Oh no, this produces the old Fibonacci. So what now? Obviously, just swap 1 and 2 to 2 and 1. Yes, this is how the new sequence that will make me famous is going to start. Now, let's see, how does it continue ...." (End of digression.)

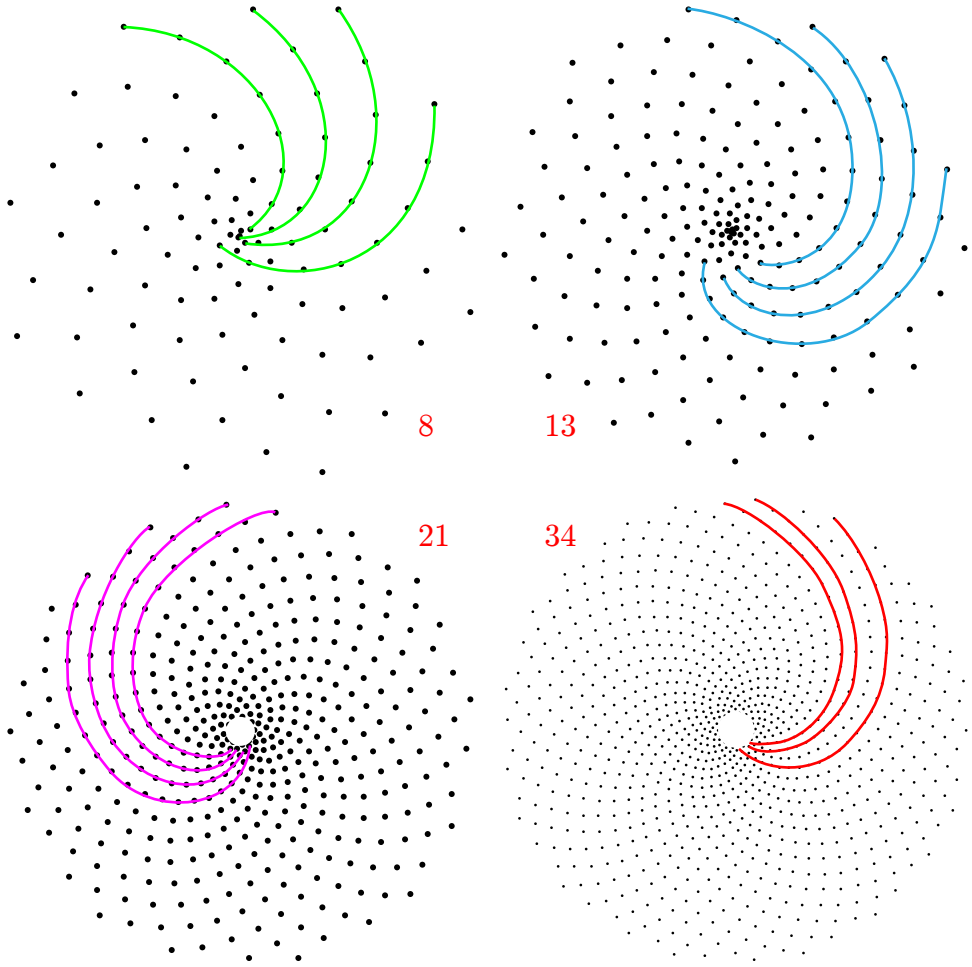
Because of the nature of the basic recursion relation, the Lucas sequence is also associated with a rectangle and spiral, (Figure 5.10) both very similar to those of the Fibonacci sequence.

At first, it seems that the Lucas sequence is irrelevant and does not contribute to our understanding of recursive sequences and their relation to φ, but it is known that this sequence is significant and is about as fundamental as the Fibonacci sequence in the following sense. If we raise φ to the nth power and then round the result off to the nearest integer, we get L_n, the nth Lucas number. Thus,

$$\text{Round}(\varphi^n) = L_n. \tag{5.2}$$

This should be compared with Equation (2.23), which relates the powers of ϕ to the Fibonacci numbers.

A surprise. Equations (2.23) and (5.2) relate the powers of ϕ to the Fibonacci numbers and Lucas numbers, respectively. The surprise is that this relation is general




```
par = GoldenRatio;
p = Table[{(n/500)Cos[2par Pi n], (n/500)Sin[2par Pi n]}, {n,0,1000}];
Graphics[{PointSize[Small], Point[p]}]
```

Figure 5.9: Four Sunflower Spirals with Fibonacci Lengths.

and exists for any sequence G_n that satisfies the recursive relation $G_n = G_{n-1} + G_{n-2}$. Reference [G.root5 18], by Matt Parker, shows that if G_n starts with the integers A and B (and continues with $A+B$, $A+2B$, $2A+3B$, etc.) then the powers ϕ^n relate to its general term G_n in the form

$$G_n = \text{Round}\left[\phi^n \frac{(3\sqrt{5}-5)A + (5-\sqrt{5})B}{10}\right]. \tag{5.3}$$

The Lucas sequence is also known as sequence A000032 in [Sloane 08]. It is related

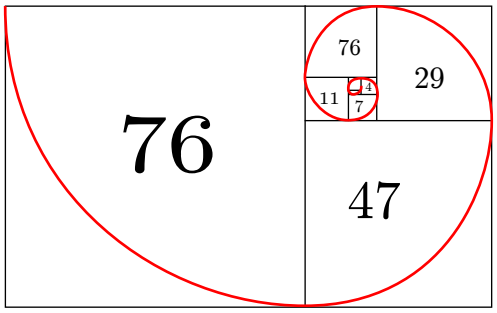


Figure 5.10: The Lucas Logarithmic Spiral.

to the Fibonacci by $L_n = F_{n-1} + F_{n+1}$ and it can be extended to negative indexes by defining $L_{-n} = (-1)^n L_n$, which results in the doubly-infinite sequence

$$\ldots, -11, 7, -4, 3, -1, 2, 1, 3, 4, 7, 11, \ldots.$$

The Lucas sequence is also found in sunflowers, where it is the second most common pattern after Fibonacci numbers.

There are also two generalized Lucas sequences, $U_n(P,Q)$ and $V_n(P,Q)$, which depend on two integer parameters P and Q and are defined by

$$U_0(P,Q) = 0,\ U_1(P,Q) = 1,\ \text{and}\ U_n(P,Q) = P\cdot U_{n-1}(P,Q) - Q\cdot U_{n-2}(P,Q)\ \text{for}\ n > 1,$$
$$V_0(P,Q) = 0,\ V_1(P,Q) = 1,\ \text{and}\ V_n(P,Q) = P\cdot V_{n-1}(P,Q) - Q\cdot V_{n-2}(P,Q)\ \text{for}\ n > 1.$$

These sequences are polynomials in P and Q, and their first seven elements are listed in Table 5.11.

n	$U_n(P,Q)$	$V_n(P,Q)$
0	0	2
1	1	P
2	P	$P^2 - 2Q$
3	$P^2 - Q$	$P^3 - 3PQ$
4	$P^3 - 2PQ$	$P^4 - 4P^2Q + 2Q^2$
5	$P^4 - 3P^2Q + Q^2$	$P^5 - 5P^3Q + 5PQ^2$
6	$P^5 - 4P^3Q + 3PQ^2$	$P^6 - 6P^4Q + 9P^2Q^2 - 2Q^3$

Table 5.11: Generalized Lucas Sequences.

5.6.2 The Plastic Ratio

The golden ratio ϕ, which is a cousin of the Fibonacci sequence, is a solution to the quadratic equation $x^2 = x + 1$. Similarly, the plastic ratio (or constant or number),

$\rho \approx 1.3247\ldots$ is the real solution of the cubic equation $x^3 = x + 1$ (its other two solutions are complex). Its exact value is

$$\rho = \sqrt[3]{\frac{9+\sqrt{69}}{18}} + \sqrt[3]{\frac{9-\sqrt{69}}{18}} \approx 1.32471795724474602596090885 4\ldots.$$

It is also easy to see that ρ satisfies the nested radical recurrence

$$\rho = \sqrt[3]{1+\sqrt[3]{1+\sqrt[3]{1+\cdots}}} = \sqrt[3]{1+\rho} \Rightarrow \rho^3 = 1+\rho.$$

(Today, we are flooded with plastic materials, but before they were invented, the term "plastic" meant something flexible or multi-purpose. One of the first to investigate the properties of this ratio was Hans van der Laan, who in 1928 named it plastic because he considered it an all-purpose ratio.)

The Padovan sequence (or Cordonnier or Perrin or van der Laan), which is named after Richard Padovan, is defined by the simple recursive relation $x_n = x_{n-2} + x_{n-3}$ and with first three elements of 1, 1, 1. Thus the sequence starts with 1, 1, 1, 2, 2, 3, 4, 5, 7, 9, 12, 16.... If we denote the first four elements by 1, x, x^2, and x^3, then the recursive relation implies the basic equation $x^3 = x + 1$, from which the rest of the sequence can be derived.

The plastic rectangle. We denote the plastic ratio by ρ and start with a rectangle of dimensions $1 \times \rho$. Partition the longer side into a long segment of $1/\rho$ and a short segment of $\rho - 1/\rho$. We continue as in Figure 5.12(b) to construct the plastic rectangle and can then draw the plastic spiral in it. This is compared to the golden equivalents, shown in part (a) of the figure (compare with Figure 2.42).

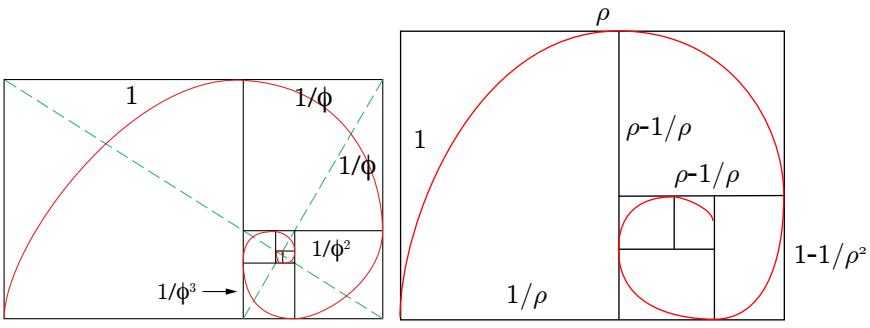
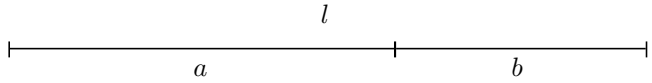


Figure 5.12: Golden and Plastic Rectangles and Spirals.

5.7 Metallic Ratios

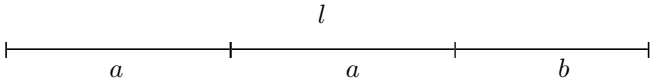
In Section 2.11.4, the golden ratio ϕ is defined by partitioning a straight segment of length l into two parts a and b such that $a > b$, $a + b = l$, and $a/b = (a + b)/a$.



This simple construction leads to the famous Fibonacci sequence (Section 5.6), as well as to the golden ratio and all its relatives. However, being mathematicians, we always try to extend concepts and go beyond what is known, so we try the following. Instead of partitioning our segment into two parts, we partition it into three parts, two equal-size large parts a followed by a short part b (i.e., $l = 2a + b$), adjust the lengths of the parts such that

$$\frac{a}{b} = \frac{2a+b}{a},$$

and name the resulting ratio the silver ratio.



It is now clear how to extend this process further. The idea is to partition the original length-l segment into n large parts a, followed by one short part b, which results in the sequence of metallic ratios

$$\begin{aligned}
\sigma_1 &\stackrel{\text{def}}{=} \frac{a}{b} = \frac{a+b}{a} = \phi,\\
\sigma_2 &\stackrel{\text{def}}{=} \frac{a}{b} = \frac{2a+b}{a},\\
\sigma_3 &\stackrel{\text{def}}{=} \frac{a}{b} = \frac{3a+b}{a},\\
&\vdots\\
\sigma_n &\stackrel{\text{def}}{=} \frac{a}{b} = \frac{na+b}{a}.
\end{aligned}$$

It turns out that many of the features of ϕ, although perhaps not all, have parallels in the other metallic ratios. The relation $\phi = 1/\phi + 1$ or $\sigma_1 = 1/\sigma_1 + 1$, can be extended to all the metallic ratios $\sigma_n = 1/\sigma_n + 1$. Similarly, $\phi^2 - \phi - 1 = 0$ can be extended to $\sigma_n^2 - n\sigma_n - 1 = 0$. The two solutions $\phi = (1 \pm \sqrt{5})/2$ can be extended to $\sigma_n = (n \pm \sqrt{n^2+4})/2$, which produces

n	σ_n	$1/\sigma_n$
1	$(1+\sqrt{5})/2 \approx 1.618033\ldots$	$0.618033\ldots$
2	$(2+\sqrt{8})/2 \approx 2.414213\ldots$	$0.414213\ldots$
3	$(3+\sqrt{13})/2 \approx 3.302775\ldots$	$0.302775\ldots$
4	$(3+\sqrt{20})/2 \approx 4.236067\ldots$	$0.236067\ldots$

The parallel to Equation (2.7) is

$$x = n + \cfrac{1}{n + \cfrac{1}{n + \cfrac{1}{n + \cfrac{1}{n + \cfrac{1}{\ddots}}}}}, \tag{5.4}$$

From which we deduct that

$$x = n + \frac{1}{x}, \text{or } x^2 - nx - 1 = 0,$$

implying that $x = \sigma_n$.

The next similarity between the golden and the general metallic ratios starts with Equation (2.7). When we truncate this continued fraction at each level and compute its value, we should not be surprised to discover that they are the famous Fibonacci numbers. Similarly, when we examine the truncated levels of Equation (5.4), we should not be surprised to learn that they are the sequence of the less famous Fibonacci-esque numbers 1, n, $n^2 + 1$, $n(n^2 + 1) + n$, ...

5.8 The Comma Sequence

The comma sequence is is based on a complex definition which causes the sequence to be open to numerous variations and extensions. The sequence is also useless; no possible practical applications of it seem possible, but this is the nature of mathematics. Its practitioners come up with all kinds of concepts, results, connections, and relations that seem worthless, but eventually prove useful and important.

The sequence depends on its first term and on the precise rule of determining the next term. Choosing 1 as the first term produces 1, 12, 35, 94, 135, 186, 248, ... Choosing 2 produces 2, 24, 71, ..., and choosing 3 brings a surprise. The entire sequence consists of 3 and 36. It simply dies after 36, since the rule prevents any number from being the next term. Here is this strange rule:

We denote the digits of the current term a_i by $\ldots abc$ and those of the (still unknown) next term a_{i+1} by $deh\ldots$. Pick the least-significant digit c of a_i and the most-significant digit d of a_{i+1}, and generate the 2-digit intermediate number cd by choosing d such that the number cd equals the difference $a_{i+1} - a_i$. The sequence's name is comma because each intermediate number straddles the comma between terms.

Starting with $a_1 = 1$, the intermediate number is $1d$ and it is easy to see that d should be 1, producing an intermediate number of 11. This causes the next term a_2 to be 12, which results in a difference of $12 - 1 = 11$, identical to the intermediate number.

Proceeding in this way, the term $a_2 = 12$ suggests an intermediate number of the form $2d$ that would produce an a_3 of the form $12 + 2d$. Again, it is easy to guess that $d = 3$ is the *smallest* candidate (although in many cases there are other candidates), leadings to $a_3 = 35$.

As is usual in today's mathematics, enthusiasts with time on their hands and fast computers at their disposal, have shown an interest in this sequence. Checking its behavior with base-10 numbers, they computed many terms and ended up with a surprise. The choice $a_1 = 1$ (sequence OEIS A330128) leads to a sequence that dies after 194,697,747,222,394 terms. Its last term is 99,999,945, which leads to an intermediate number $5d$. The next term can therefore start only with 9 or with 1, but $99999945 + 59 = 100000004$ and $99999945 + 51 = 99999996$, both with wrong differences.

Another surprise is that choosing $a_1 = 3$ ends after only two terms, namely 3 and 36. The reader may find it easy to verify that there is no way to add an intermediate number of the form $x = 6d$ to 36 and ends with a difference of $x - 36 = 3d$ for any decimal digit d. A surprise, a quick death, and a strange sequence indeed!

Research on the comma sequence has so far suggested that there are certain numbers that kill it, and that they tend to cluster. Examples are 18, 27, 36, 45, 54, 63, 72, and 81; 918, 927, 936, 945, 954, 963, 972, and 981; 9918, 9927, 9936, 9945, 9954, 9963, 9972, and 9981. As the terms of a comma sequence get bigger, the clusters become more and more dispersed, but research in this area is still in its infancy, and there is no certainty that it will continue.

We end with some ideas about this sequence:

- Commas are crucial in the comma sequence, but they are used in two places, inside the digits of large terms and between terms. Thus, it's a good idea to use perhaps semicolons to separate the terms of the sequence.

- It is possible, even easy, to come up with extensions and variations of this sequence. How about a sequence whose general term a_n is the total length of the comma sequence that starts with n.

- A multiplicative comma sequence can be defined such that the next term is the product (instead of the sum) of the current term and their intermediate number. In the discussion above, the next term is determined in each set by choosing the smallest candidate. A possible extension is to check all feasible candidates and choose the one that results in the longest sequence. A simple case is the choice $a_1 = 2$, which leads to an intermediate number $2d$ and the choice $d = 4$ or $d = 5$. The latter choice results in the short multiplicative sequence 2 and $2 \cdot 25 = 50$, whereas the former choice results in the two branches 2, $2 \cdot 24 = 48$, and 4032; and 2, $2 \cdot 24 = 48$, 3984, 163344, 7677168, 660236448, and 56120098080.

◇ **Exercise 5.10:** The following numeric sequences, discussed in [funny.sequences 18], are defined by simple rules that non-mathematicians may find misleading or even childish. Try to guess them.

1, 4, 8, 48, 88, 488, 888, 4888, 8888, ...
61, 21, 82, 43, 03, 6....
0, 0, 0, 0, 9, 5, 1, 1, 0, 55, ...
2, 4, 6, 30, 32, 34, 36, 40, 42, 44, 46, 50, 52, 54, 56, 60, 62, 64, 66, 2000, ...
13, 17, 31, 37, 71, ...
A bonus sequence: 131, 228, 331, xxx, 531, 630,...
(End of exercise.)

5.9 Quasi-Numeric Sequences

The elements of a sequence may be numeric, but the rule determining the next element may be nonnumeric. Such a sequence may be referred to as quasi-numeric. The following elegant "look and say" sequence sequence, one of the many brainchilds of John H. Conway, elegantly illustrates this concept (see also Page 69).

1	one
11	one one
21	two ones
1211	one two and one one
111221	one one, one two, two ones
...?...	three ones, two twos, one one
	one three, one one, two twos, two ones

It may be easy to come up with such a sequence, but extremely difficult to guess its rule. Even a very simple example of a quasi-numeric sequence, such as the sequence `m`, `t`, `w`, `t`, `f`, and `s`, is confusing (these are the first letters of the names of the days of the week in English, starting with Monday). Here are a few examples.

1. Which two integers come next in the infinite sequence 38, 24, 62, 12, 74, ...?

Answer. The next two integers are 28 and 102. The rule is simple but elusive. Start with (almost) any positive 2-digit integer (we somewhat arbitrarily selected 38). Multiply the two digits to obtain $3 \times 8 = 24$, then add $38 + 24$ to generate the third integer 62. Now multiply $6 \times 2 = 12$ and add $62 + 12 = 74$. Similar multiplication and addition produce the next two integers 28 and 102.

2. What is the next integer x in the sequence $(12, 6)$, $(6, 3)$, $(10, x)$?

Answer. The next integer is 3. The first integer of each pair is random, and the second one is the number of letters in the English name of the integer.

3. What is the next integer in the sequence 3, 7, 10, 11, 12, ...?

Answer. It is 17. The elements of this sequence are those integers whose names, when fully spelled in English, contain `E` but not any other vowels.

4. What is the next integer in the sequence 5, 25, 61, 113, 181, ...?

Answer. It is 265. Each integer is the sum of two consecutive squares. Thus, $1^2 + 2^2 = 5$, $3^2 + 4^2 = 25$, and so on.

5. What are the next two integers in the sequence? 2, 4, 6, 30, 32, 34, 36, 40, 42, 44, 46, 50, 52, 54, 56, 60, 62, 64, and 66?

Answer. This is the sequence of all the integers whose names, when fully spelled in English, do not contain the vowel `E`. All the integers of the form $7x$, $8x$, $9x$, $1xy$, $2xy$, and so on until 2000 have an `E` in their names. Thus, the next two integers are 2000 and 2002. This is the Eban (for `E` is banned) sequence, also known as sequence `A006933` in [Sloane 08].

6. What is the rule of the sequence 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 6, 6, 6, 6, 6, 9, 9, 9, 9, 9, 13,...?

Answer. The nth element of this sequence is the number of ways of making change for n cents using coins of 1, 5, 10, and 25 cents.

7. What is the rule of the sequence 1, 2, 3, 2, 1, 2, 3, 4, 2, 1, 2, 3, 4,...?

Answer. The nth element of this sequence is the number of letters in the Roman numbering of the integer n.

8. What is the rule of the sequence 1, 2, 4, 8, 16, 77, 145, 668, 1345, 6677, 13444, 55778,...?

Answer. Given 145, we reverse it to obtain 541, add $145 + 541 = 686$ and sort the digits to obtain the next element 668. This is the RATS sequence (for reverse, add, then sort), number `A004000` in [Sloane 08].

A fun sequence. The elements of the sequence 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 20, 20, 100, and 1000 are all the numbers whose names in French are monosyllabic (un, deux, trois, quatre, cinq, six, sept, huit, neuf, dix, onze, douze, treize, quinze, seize, vingt, trente, cent, mille.)

5.10 Series

In english, the term "series" is both singular and plural. In mathematics, a series is the sum of the elements (or terms) of a sequence. The series that are of interest in mathematics are infinite, and the main feature of an infinite series is whether it converges and to what limit. A series $\sum_i^\infty a_i$, with real or complex elements a_i, converges if the sequence S_n of its partial sums $S_n = \sum_i^n a_i$ converges to a finite limit as n grows without limit. However, because of the enigmatic nature of infinity, infinite series can feature various types of convergence and can be confusing as our first example illustrates. The example involves the alternating harmonic series

$$S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} + \frac{1}{11} - \cdots. \tag{5.5}$$

All we need to do is multiply both sides by 2

$$2S = 2 - 1 + \frac{2}{3} - \frac{1}{2} + \frac{2}{5} - \frac{1}{3} + \frac{2}{7} - \frac{1}{4} + \frac{2}{9} - \frac{1}{5} + \frac{2}{11} - \cdots,$$

and collect terms with the same denominator

$$2S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots. \tag{5.6}$$

comparing Equations (5.5) and (5.6) suggests that $2S = S$ or $2 = 1$, and implies that the problem of convergence of infinite series must be taken seriously and dealt with carefully.

Years of research have resulted in the following classification of infinite series according to how they converge:

- Absolute convergence. A series is said to converge absolutely (or to be absolutely convergent) if the sum $\sum_i |a_i|$ of the absolute values of its elements is finite. The reason for the term "absolute" is that any rearrangement or reordering of the elements of the series will still converge to the same value. An example is the Maclaurin series of the

exponential function $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$. Another example is the alternating sum of the reciprocals of the square integers

$$1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \frac{1}{25} - \frac{1}{36} + \frac{1}{49} - \frac{1}{64} + \frac{1}{81} - \cdots$$

which converges to 0.822467.... This convergence is absolute, because the similar series of the absolute values of those reciprocals also converges (to 1.644934...).

- Conditional convergence. A series is said to be conditionally convergent if it converges, but it does not converge absolutely. In other words, if the infinite sum of its elements has a finite limit, but the infinite sum of the absolute values of the same elements diverges. An example is the harmonic series, $\sum_{n=1}^{\infty} \frac{1}{n}$, whose elements are all positive. This series diverges (albeit very slowly), but its alternating version, with alternating positive and negative values, converges to ln(2), and is therefore conditionally convergent.

  Another example is the series of alternating reciprocal odd numbers

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \frac{1}{13} - \frac{1}{15} + \frac{1}{17} - \cdots$$

  which converges to $\pi/4$, but the series of the absolute values of its terms diverges (also very slowly).

  An unexpected result of this definition is that the elements of a conditionally convergent series may be rearranged to converge to any given real value. They may even be rearranged to diverge to ∞ or to $-\infty$.

- Unconditional convergence. This term refers to a series where any rearrangement of its elements converges to the same value. Every series which converges absolutely also converges unconditionally.

- Uniform convergence. Let $\big(f_1(x), f_2(x), f_3(x), \ldots\big)$ be a sequence of functions. The series $\sum_{n=1}^{\infty} f_n(x)$ is said to converge uniformly to f if the sequence of partial sums

$$S_n(x) = \sum_{k=1}^{n} f_k(x)$$

converges uniformly to f.

Uniform convergence of functions is defined as follows: Given any arbitrarily small positive number ε, a number N can be found such that each of the functions f_N, f_{N+1}, f_{N+2}, ... differs from the limit f by no more than ε at every point x.

The Riemann series theorem

Now that the various types of convergence have been introduced, we are ready for the surprising and unintuitive Riemann series (or Riemann rearrangement) theorem. This theorem applies only to conditionally convergent infinite series whose terms are real numbers. Given an arbitrary real number α and such a series, its terms can be rearranged (permuted) so that the new series converges to α. Even more, it is always possible to rearrange the elements so that the new series will diverge to ∞ or to $-\infty$.

This theorem also implies that a series of real numbers is absolutely convergent if and only if it is unconditionally convergent. A series that is absolutely convergent is not affected by rearranging, and always converges to the same limit.

The alternating harmonic series, Equation (5.5), is an example. We know that it converges to ln(2), Now rearrange it to

$$1-\frac{1}{2}-\frac{1}{4}+\frac{1}{3}-\frac{1}{6}-\frac{1}{8}+\frac{1}{5}-\frac{1}{10}-\frac{1}{12}+\frac{1}{7}-\cdots,$$

which can be broken into the sum of triplets

$$-\frac{1}{4k}-\frac{1}{4k-2}+\frac{1}{2k-1}=\frac{1}{2(2k-1)}-\frac{1}{2(2k)},$$

allowing us to write it as

$$\frac{1}{2}-\frac{1}{4}+\frac{1}{6}-\frac{1}{8}+\frac{1}{10}-\frac{1}{12}=\frac{1}{2}\left(1-\frac{1}{2}+\frac{1}{3}-\cdots\right),$$

which equals one-half the original alternating harmonic series, and converges to ln(2)/2.

While the Riemann series theorem is a surprise, it is easy to explain and understand. A conditionally convergent series must contain both positive and negative terms, and must have infinitely many of each type. If there is only a finite number of negative terms, then they must add up to a finite number L, while the positive number would sum without limit.

Here is a simple proof of this theorem. Imagine that we want to rearrange the terms of a conditionally convergent series so they converge to the real number α. We first sort the positive terms in order from the largest to the smallest, and sort the negative terms similarly. We then start adding positive terms from the largest one until the sum exceeds α, and follow by subtracting negative terms (again using the largest ones) until the combined sum drops below α. This process continues, as the sum oscillates around α, but the oscillations get smaller, because we use smaller and smaller terms.

The following notes cover a few subtle points that may be hidden in the above proof.

- The proof implicitly assumes that the series has two largest elements, one positive and the other negative. In principle, an infinite series may contain elements that get larger without limit, but the elements of a convergent series, absolute or conditional, cannot be large without a limit. Similarly, there cannot be a smallest element. If a series converges, its elements must get smaller and approach zero without a limit. Also, any subsequence must eventually become small, which is why there cannot be large elements hiding among the smaller elements.

- The proof assumes that a conditionally convergent series must contain both infinite positive and infinite negative terms. Consider a conditionally convergent series that converges to a limit L. If this series had, say, a finite number of negative terms, then changing their signs would not make the series diverge. It would only change its limit to $L+$(twice the sum of the negative elements). Thus the two lists of positive and negative terms must be infinite and each must get smaller and smaller without limit.

- Is it possible that the same-sign terms added in one of the steps above would never reach α? After all, the terms that we add get smaller and smaller and may reach a limit before they reach α. A good example is the familiar geometric series

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\cdots,$$

where each term being added covers half the remaining distance to 2. The term increases the sum slightly, and yet the number 2 is the limit of the infinite sum, and the sum never actually reaches this limit.

The answer is that in a conditionally convergent series, the set of positive terms must be divergent, and the same is true for the set of negative terms. This implies that there would always be enough terms to reach any α and beyond.

- Another potential problem to consider is a possible lack of small terms. Is it possible that all the terms of the series would be large, thereby causing the process above to oscillate about α without ever converging to it? This also is impossible, because for the entire series to converge (conditionally or absolutely), it must have terms that get smaller without limit.

Convergence of series

Here is an example that illustrates the baffling nature of convergence. The harmonic series $1+1/2+1/3+1/4+1/5+\cdots$ diverges, albeit very slowly. However, if we omit certain terms, the modified series may converge to a finite sum. For example, if we omit the terms $1/n$ for all n whose decimal value contains at least one zero (such as 10, 50, 103, and 200), then the sum of the remaining terms converges to 23.10345....

The P-series. The harmonic series is only a special case of a p-series, which itself is a special case of the Riemann zeta function $\zeta(s)$ for real values of s. P-series is a family of infinite series defined by

$$\sum_{n=1}^{\infty}\frac{1}{n^p}=1+\frac{1}{2^p}+\frac{1}{3^p}+\cdots.$$

The p-series converges for values $p>1$ and diverges for any $p\leq 1$. This is known as the p-test, which is proved here.

◇ **Exercise 5.11:** Apply the p-test to the series $\sum_{n=1}^{\infty}\frac{n}{n^{3/2}+3}$.

The proof of the p-test is straightforward and requires no big guns. We start with the simple inequality

$$\begin{aligned}0<\int_1^{\infty}\frac{1}{x^p}\,dx<\sum_{n=1}^{\infty}\frac{1}{n^p}<1+\int_1^{\infty}\frac{1}{x^p}\,dx\\ \stackrel{\text{def}}{=}0<A<\sum_{n=1}^{\infty}\frac{1}{n^p}<B.\end{aligned}\tag{5.7}$$

It is obvious that if integral A diverges, the infinite p-series sum of Equation (5.7) diverges too. Similarly, if B converges, then the sum also converges. On the other hand,

if the sum converges, then A will converge, and if the sum diverges, B too will diverge. Thus, the p-series of Equation (5.7) converges if and only if the integral of A and B converges.

We therefore concentrate on this integral and start by deriving its indefinite form, which is $x^{1-p}/1-p$. Because the denominator is undefined for $p=1$, we must treat this case separately and we rewrite integral A as the limit

$$\int_1^\infty \frac{1}{x^p}\,dx = \lim_{k\to\infty}\int_1^k x^{-p}\,dx.$$

For $p=1$, the integral becomes $\ln(k)-\ln(1)$, which diverges at the limit. For $p\neq 1$, we obtain

$$\lim_{k\to\infty}\int_1^k x^{-p}\,dx = \lim_{k\to\infty}\frac{x^{1-p}}{1-p}\Big|_1^k = \lim_{k\to\infty}\left[\frac{k^{1-p}}{1-p}-\frac{1^{1-p}}{1-p}\right],$$

which is the same as the limit of k^{1-p} for large values of p.

It is now clear that for $p>1$, where $1-p$ is negative, the term k^{1-p} is getting smaller with large values of k, and its limit is therefore zero, implying that integral A of Equation (5.7) converges. Similarly, for $p<1$, $1-p$ is positive, the limit of k^{1-p} is ∞, and integral A of Equation (5.7) diverges. This proves the validity of the p-series test.

For most values of p, the sum of a p-series series is unknown. Here are a few known examples:

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6},\quad \sum_{n=1}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{90},\quad \sum_{n=1}^{\infty}\frac{1}{n^6}=\frac{\pi^6}{945}.$$

A non-trivial example of a quasi P series. The famous Euler number e can be defined as the limit of the interesting expression $(1+1/n)^n$ for large n. When this is turned on its head, it becomes $n^{1+1/n}$, and the above discussion of P series makes it natural to ask about the behavior of the series

$$\sum_{n=1}^{\infty}\frac{1}{n^{1+1/n}}. \tag{5.8}$$

This series is similar to the various P series, except that its exponent $1+1/n$ is not fixed but goes down from 2 to 1, implying that the general term of the sequence goes down from 1 to zero.

The proof presented here is not the most elegant, but is straightforward. We start by showing that for all integers n, it is true that $n^{1/n}<2$, and therefore $1/n^{1/n}>1/2$. This is based on the well-known binomial expansion theorem, first derived by Newton.

$$2^n=(1+1)^n=(x+1)^n|_{x=1}=\sum_{k=0}^{n}\binom{n}{k}x^k\Big|_{x=1}=1+n+\frac{n(n+1)}{2}+\cdots>n,$$

which implies $2^n>n$ and therefore $2>n^{1/n}$. Based on this, we can rewrite Equation (5.8)

$$\sum_{n=1}^{\infty}\frac{1}{n^{1+1/n}}=\sum_{n=1}^{\infty}\frac{1}{n^{1/n}}\frac{1}{n}>\sum_{n=1}^{\infty}\frac{1}{2}\frac{1}{n}.$$

This shows that the infinite sum of Equation (5.8) is greater than half the harmonic series, and it therefore diverges.

5.10.1 The Harmonic Series

.

The infinite sum of reciprocals

$$\sum_{n=1}^{\infty}\frac{1}{n}=1+\left[\frac{1}{2}\right]+\left[\frac{1}{3}+\frac{1}{4}\right]+\left[\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right]+\left[\frac{1}{8}+\frac{1}{9}+\cdots\right],$$

is called the harmonic series. The name harmonic comes from music, where the harmonics, or overtones, created by a vibrating string feature wavelengths that are multiples of 1/2, 1/3, 1/4, etc., of the basic wavelength.

The harmonic series is perhaps the simplest infinite series that has a chance of converging to a finite value, since its elements get smaller and smaller. However, it is easy to show that it diverges (albeit very slowly), and this has already been done by the 14th century. Consider the groups of elements in square brackets. The second group consists of two elements, both greater than or equal 1/4, so this group is greater than 1/2. The third group consists of four elements, each greater than or equal 1/8, which is why this group is greater than $4(1/8) = 1/2$, and so on.

The harmonic series diverges, but extremely slowly. It takes 12,367 terms for it to exceed 10, 272,400,600 terms to exceed 20, and (approximately) $1.5 \cdot 10^{43}$ terms to exceed 100.

Partial sums of the harmonic series are called harmonic numbers and are denoted by H_n. Thus

$$H_n=\sum_{k=1}^{n}\frac{1}{k}.$$

5.11 The Real Harmonic Series

The harmonic series $\sum^{\infty}(1/k)$ is defined for integer values of k, which is why it was only natural that mathematicians, being curious, would try to extend it to any real argument x. This section, following the clear presentation of reference [Lines That Connect 21], explains how the discrete sequence of values that make up the harmonic series can be generalized to a continuous, real function.

A simple example of extending a sequence of values to a continuous function is the sum

$$\sum_{k=1}^{n}k$$

of the first n natural numbers. This series is defined only for integers k, but is easy to extend to any real numbers, because it is trivial to prove (e.g., by induction) that this sum equals $n(n+1)/2$.

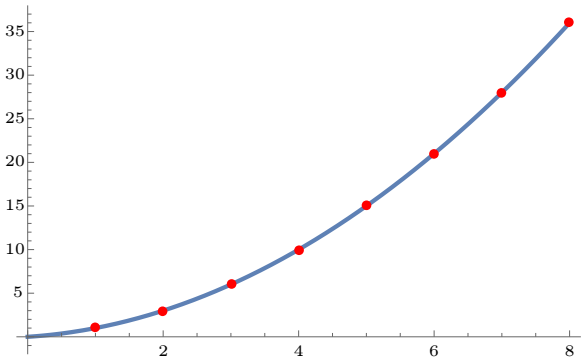
Now, that n is not limited to integer values and can have any real value, we can have, for example, the sum

$$\sum_{k=1}^{5.7} k = \frac{5.7(5.7+1)}{2} = 19.095,$$

which is somewhere between $\sum^5 k = 15$ and $\sum^6 k = 21$. Figure 5.13 shows how the real function

$$f(x) = \frac{x(x+1)}{2},$$

is graphed as a continuous, smooth curve (part of a parabola) that passes through the points of the discrete sum.




```
n=8;
L2=ListPlot[Table[{k,Sum[i,{i,1,k}]},{k,1,n}],
 PlotStyle->Directive[AbsolutePointSize[6],Red]];
L1=Plot[x(x+1)/2, {x,0,n}];
Show[{L1,L2}]
```

Figure 5.13: Discrete and Continuous Sums of n Integers.

Other sums, such as

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \quad \text{and} \quad \sum_{k=1}^{n} 2^k = 2(2^n - 1),$$

can be extended from discrete integers to continuous reals, and this section explains how to achieve this type of generalization with the harmonic series. The first step is to denote the partial sums of this series by

$$H(n) = \sum_{k=1}^{n} \frac{1}{k}, \tag{5.9}$$

which straight away produces the three related recursive relations

$$H(n) = H(n-1) + \frac{1}{n}, \tag{5.10}$$

$$H(n+1) = H(n) + \frac{1}{n+1}, \tag{5.11}$$

$$H(n-1) = H(n) - \frac{1}{n}. \tag{5.12}$$

Notice that Equation (5.12) allows us to compute $H(0) = H(1) - \frac{1}{1} = 0$, but when we try to extend this further and compute $H(-1) = H(0) - \frac{1}{0} =$ undefined, we must give up and realize that the harmonic series does not extend naturally into the negative integers.

The next step is to write Equation (5.11) in terms of real numbers in the form $H(x+1) = H(x) + \frac{1}{x+1}$, and extend it from $x+1$ to $x+n$, where it has the form

$$H(x+n) = H(x) + \sum_{k=1}^{n} \frac{1}{x+k}. \tag{5.13}$$

The resulting Equation (5.13) will become one of the two components needed to produce the final real harmonic series.

The other component is an interpolation that would propagate our final curve smoothly and continuously, segment by segment, from any point $H(n)$ on the curve to its successor $H(n+1)$, such that individual segments connect smoothly to form one shiny, unwrinkled curve. We know that the terms of the harmonic series get smaller and smaller, which causes their partial sums to get closer and closer. The limit of the difference of consecutive partial sums is therefore zero. Thus,

$$\lim_{N\to\infty} (N+x) - H(N) = 0,$$

and this limit can be employed to produce the desired result. We simply substitute Equation (5.13) for the first term and Equation (5.9) for the second term, to obtain

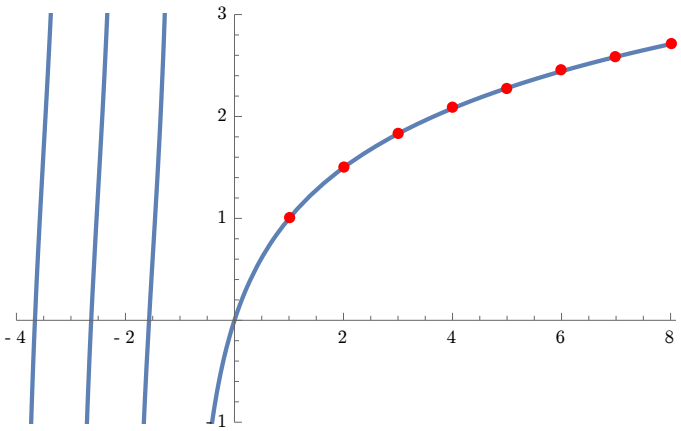
$$\begin{aligned}
&\lim_{N\to\infty} (N+x) - H(N) \\
=&\lim_{N\to\infty} H(x) + \sum_{k=1}^{N} \frac{1}{x+k} - \sum_{k=1}^{N} \frac{1}{k} \\
=&H(x) + \lim_{N\to\infty} \sum_{k=1}^{N} \frac{1}{x+k} - \sum_{k=1}^{N} \frac{1}{k}.
\end{aligned}$$

which implies

$$\begin{aligned}
H(x) &= \lim_{N\to\infty} -\sum_{k=1}^{N} \frac{1}{x+k} + \sum_{k=1}^{N} \frac{1}{k} \\
&= \lim_{N\to\infty} \sum_{k=1}^{N} \left[\frac{1}{k} - \frac{1}{x+k}\right].
\end{aligned}$$

Finally, the expression for the real harmonic series is obtained by replacing the limit with an infinite sum

$$H(x) = \sum_{k=1}^{\infty} \left[\frac{1}{k} - \frac{1}{x+k} \right].$$




```
n=8;
L2=ListPlot[Table[Sum[1/i,{i,1,k}],{k,1,n}],
 PlotStyle->Directive[AbsolutePointSize[6],Red]];
L1=Plot[Sum[(1/k)-(1/(x+k)), {k,1,1000}], {x,-4,n},PlotRange->{{-4,8.1},{-1,3}}];
Show[{L1,L2}]
```

Figure 5.14: Discrete and Continuous Harmonic series.

Figure 5.14 shows some points of the harmonic series, together with its continuous curve. A part of the negative x-axis is also included, and it shows the strange, unexpected behavior of our curve. It turns out that this curve is a shifted version of the less-known digamma function $\psi(x)$, which is the logarithmic derivative of the gamma (Γ) function, Section 10.1.

5.12 The Book-Stacking Problem

Reference [WikiHarmonic 23] describes several applications of the harmonic series, perhaps the most well-known of which is the block-stacking (or book-stacking) problem. Given an unlimited number of rectangular books or blocks that are identical, uniform, and have a length of one unit and a weight of one unit, how far away can they be stacked vertically and overhang over the edge of a table without falling.

The key to understanding this problem, following reference [StackBlocks 16], is the physics of center-of-gravity or moment. Imagine a uniform block partly protruding over the edge of a table. The difference

$$(\text{mass} \times \text{length of part on the table}) - (\text{mass} \times \text{length of part protruding})$$

is the difference of its moments. The block can be slid toward the edge without dropping until this difference is zero. The same difference of moments also governs the behavior of a set of blocks.



Figure 5.15: Image by MS Copilot, Sept. 2025.

We use induction to determine the difference of moments of a set of blocks. This method starts with a set of $n-1$ blocks that are already balanced on the edge of a table. An nth block is then slid under the set and the problem is to figure how it affects the position of the total center of gravity. Figure 5.16 illustrates a set of $n-1=3$ blocks with its center of mass (red in the figure) right at the edge of a table. We now imagine that the blocks are lifted up and an nth (fourth) block is inserted under them with its front flush with the edge of the table. The new block is to the left of all the other blocks and has shifted the center of gravity of the set of n blocks to the left (indicated in green in the figure). We denote the distance between the old and new centers of gravity by x.

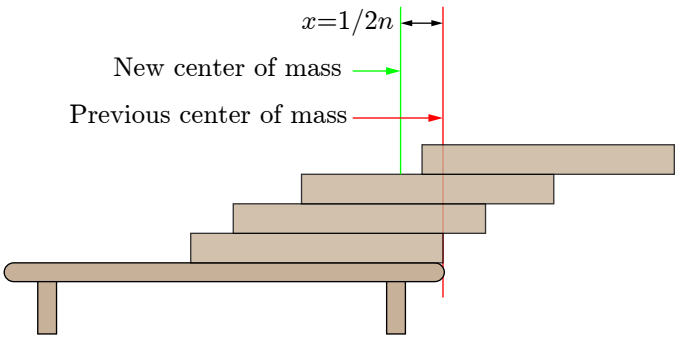


Figure 5.16: Four Stacked Blocks.

On one hand, the center of gravity of new block is 1/2 a unit from the edge of the table, so its moment is the product of its mass (which is 1) and half its length (which is 1/2) or 1/2. On the other hand, the new set of n blocks, which used to have a zero

difference of moments, now has a net difference of moments of xn toward the left. This difference is due to the new block, which allows us to write $xn = 1/2$ or $x = 1/(2n)$.

Thus, we can now slide the new set of blocks a distance $1/(2n)$ to the left without it dropping down. To complete the induction we notice that for $n = 1$, this distance is $1/2$, which makes sense since half of the first block can safely protrude past the edge.

For five blocks, the relation $x = 1/(2n)$ produces the series of lengths

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\frac{1}{10}=\frac{1}{2}\left[1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}\right]=\frac{137}{120},$$

which is the harmonic series.

In theory, given enough blocks, a set of blocks can protrude any distance beyond the edge of a table. In practice, however, non uniformity and minute variations in mass and length put a limit on what can be achieved.

> In theory there is no difference between theory and practice, but in practice there is.
> —Yogi Berra.

The following series comes up in one of the historical paradoxes by Zeno of Elea (see him again in Section 6.2). Here, we discuss just the series, which is

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots, \tag{5.14}$$

which converges to 2.

◇ **Exercise 5.12:** Prove that the series of Equation (5.14) converges to 2.

Euler series. The next natural question at this point is how to modify the harmonic series so it converges, and the simple modification suggested here does just that. The harmonic series adds the reciprocals, which are less than 1, so how about adding their squares, which are smaller? It is easy to prove the convergence of the series

$$\sum_{n=1}^{\infty}\left(\frac{1}{n}\right)^2=1+\left(\frac{1}{2}\right)^2+\left(\frac{1}{3}\right)^2+\left(\frac{1}{4}\right)^2+\left(\frac{1}{5}\right)^2+\left(\frac{1}{6}\right)^2+\left(\frac{1}{7}\right)^2+\cdots. \tag{5.15}$$

The second and third terms are each smaller or equal $1/4$, so their sum is smaller than $1/2$. The next four terms are each smaller or equal $1/16$, so their sum is smaller than $4\cdot 1/16 = 1/4$, and so on. Each term in this series is smaller than the corresponding term in the well-known converging series of Equation (5.14), which shows that series (5.15) also converges, but to what value? It was Euler (who else) who proved that its sum is $\pi^2/6 \approx 1.644$, which is why this series is named after him. (An alternative name is the Basel problem.)

The Euler proof presented here follows the description in reference [Dunham 91]. It is ingenious and unexpected, because it starts with the familiar sine function, which comes from trigonometry and seems to have no connection with the Euler series. The graph of $\sin x$ is a wave which cuts across the x-axis at $x = 0, \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$. These values are the zeros of sine.

Once the zeros of the sine function are understood, we switch to the Taylor series of sine

$$\sin x = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots,$$

which expresses sine as a polynomial with only odd powers of x. Like any other polynomial, this can be expressed as the product of terms that depend on its roots. In general, given a polynomial $P(x)$ that satisfies $P(0) = 1$, and with roots r_i that are nonzero, it can be written in the form

$$P(x) = \left[1 - \frac{x}{r_1}\right]\left[1 - \frac{x}{r_2}\right]\left[1 - \frac{x}{r_3}\right]\cdots. \tag{5.16}$$

It is easy to see that this satisfies $P(r_i) = 0$ and $P(0) = 1$.

The next step is to divide this Taylor series by x (and remember that from now on, x cannot be zero).

$$\frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \frac{x^8}{9!} - \cdots. \tag{5.17}$$

The roots of this polynomial are the same as the roots of $\sin x$, except that we must ignore the root $x = 0$. Once we substitute those roots as in Equation (5.16), we end up with

$$\frac{\sin x}{x} = \left[1 - \frac{x}{\pi}\right]\left[1 - \frac{x}{-\pi}\right]\left[1 - \frac{x}{2\pi}\right]\left[1 - \frac{x}{-2\pi}\right]\left[1 - \frac{x}{3\pi}\right]\left[1 - \frac{x}{-3\pi}\right]\cdots. \tag{5.18}$$

Combining Equations (5.17) and (5.18) and applying the identity $(a+b)(a-b) = a^2 - b^2$, we obtain our key equation

$$1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \frac{x^8}{9!} - \cdots = \left[1 - \frac{x^2}{\pi^2}\right]\left[1 - \frac{x^2}{4\pi^2}\right]\left[1 - \frac{x^2}{9\pi^2}\right]\left[1 - \frac{x^2}{16\pi^2}\right]\cdots.$$

In addition to bringing us very close to the elusive proof, this equation is ingenious because it equates an infinite sum to an infinite product.

The almost-last step is to multiply out this infinite product, which happens to be easy because of its special form. The result is

$$1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \frac{x^8}{9!} - \cdots = 1 - \left[\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \frac{1}{16\pi^2} + \cdots\right]x^2 + (\ldots)x^4 - \cdots.$$

And the last step is to equate the coefficients of x^2 on both sides, which yields

$$-\frac{1}{3!} = -\left[\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \frac{1}{16\pi^2} + \cdots\right],$$

or

$$\frac{1}{6} = \frac{1}{\pi^2}\left[1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots\right],$$

An expression that leads directly to the final value $\pi^2/6$ as any reader should be able to verify.

Once Euler achieved this breakthrough, he employed his method, using $\sin x/x$ for nonnegative x, to solve related problems, such as the sum of reciprocals of even squares $\sum 1/(2k)^2$, the sum of odd perfect squares $\sum 1/(2k+1)^2$, and the sum $\sum 1/k^4$.

The Euler series is a special case of the all-important Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

And Euler computed its values for all even integers. Thus, $\zeta(2) = \pi^2/6$, $\zeta(4) = \pi^4/90$, and $\zeta(6) = \pi^6/945$. The values of the zeta function for odd integers are still unknown, but Euler came up with the baffling expression

$$\zeta(-1) = 1 + 2 + 3 + \cdots = \frac{-1}{12},$$

a result about which people argue endlessly on YouTube and elsewhere.

[Numbers] The difference between hearsay and prophecy is often one of sequence. Hearsay often turns out to have been prophecy.

—Hubert H. Humphrey*(1985)*.



6
Paradoxes

The term paradox is frequently heard. We also learn about paradoxes and their resolutions or lack thereof, but what precisely is a paradox? A paradox is a puzzling issue or riddle with a solution that appears to be unattainable. Certain paradoxes appear to challenge our intuitive understanding or to reject our familiar grasp on reality. Paradoxes can be created for amusement or education purposes, yet they can also represent contradictions between our perceptions and beliefs.

The term paradox, from the Greek $\Pi\alpha\rho\alpha\delta o\xi o$, is the concatenation of para ($\Pi\alpha\rho\alpha$, meaning "opposite" or "distinct from") and $\Delta o\xi\alpha$ (Doxa, "I express opinion or knowledge"). So a paradox is something that is opposite to my opinion or distinct from our knowledge.

6.1 Types of Paradoxes

Like many objects that we see around us, ideas that we might have, and terms that we use, we intuitively try to distinguish different types of paradoxes; we try to classify them in an attempt to better understand them. The following three categories of paradoxes are due to the philosopher Willard Quine, who defined them in 1961.

- A falsidical paradox packs a surprise, but it is seen as a false alarm when we solve the underlying fallacy. Such a paradox is a statement or question that appears to be logical at first but upon closer examination leads to an impossibility. It can also be described as a problem that uses an incorrect assumption to justify a result that is false. A simple example is Which came first, the chicken or the egg?

- A veridical paradox packs a surprise, but the surprise quickly dissipates itself as we ponder the proof. Such a paradox is a statement that seems contradictory or even false but is actually true. It can also be described as a situation that is ultimately, logically

true, but is either senseless or ridiculous. A simple example is the phrase The more things change, the more they stay the same (Plus ça change, plus c'est la même chose).

These two types have solutions and are paradoxes only to those who don't know the solutions.

■ An antinomy (literally, against law) packs a surprise that can be accommodated by nothing less than a repudiation of part of our conceptual heritage. This type is a paradox to all of us, since it has no solution. However, advances in science or mathematics can provide a solution, which is why an antinomy (not to be confused with antimony, which is chemical element of atomic number 51) may, in future, be resolved and become a paradox of the first two types. As Quine said "One Man's antinomy is another man's falsidical paradox, give or take a couple thousand years."

An antinomy can also be described as a statement that lays out a set of conditions and then asks a question, the resolution of which becomes self-contradictory, resulting in lack of a valid answer.

◇ **Exercise 6.1:** In his book *Life of Theseus*, Plutarch tells us "The ship wherein Theseus and the youth of Athens returned [from Minoan Crete] had thirty oars, and was preserved by the Athenians down even to the time of Demetrius Phalereus, for they took away the old planks as they decayed, putting in new and stronger timber in their place . . . ." This story is the origin of the well-known Ship of Theseus paradox. Consider the following question, is the resulting ship the same as the original ship? In general, this is equivalent to the question what makes something truly original? An alternative version has to do with the body of an animal, where new cells are constantly replacing old ones. What type of paradox is this and what can we say about its resolution?



Figure 6.1: Image by MS Copilot, Sept. 2025.

◇ **Exercise 6.2:** Come up with a paradox or a question similar to the Ship of Theseus paradox.

■ Some philosophers suggest a fourth type of paradox, a dialetheia. This is similar to an antinomy, but states that both a statement and its opposite can both be true

simultaneously. I could not find examples of this type, but there are arguments for why they should or should not exist. It is a question of finding a condition that can be true simultaneously with its opposite.

> It is best to start by explaining what is meant by a paradox. The word is used in different ways, of course, but at the heart of all the uses is contradiction. A paradox starts with a set of reasonable premises. From these premises, it deduces a conclusion that undermines the premises. It is a travesty of the notion of proof.
> One thing not immediately apparent (if the paradox is clever enough) is the reason for the contradiction. Is it possible for a perfectly valid argument to lead to contradiction, or is that "guaranteed" not to happen?
>
> —William Poundstone, *Labyrinths of Reason*, 2011.

6.2 Examples of Paradoxes

• An example of a falsidical paradox is the racecourse, or dichotomy paradox, by Zeno of Elea. Suppose a runner needs to travel from a start `S` to a finish `F`. To do this he must first travel to the midpoint, `M`, and thence to `F`, but in order to reach `M`, he must first get to `N`, the midpoint of `SM`. Extending this argument ad infinitum proves that the runner cannot take a single step, let alone finish his run, because it is impossible to accomplish an infinite number of tasks in a finite time.

This paradox started as an antinomy. It must have baffled the ancient Greeks because no one knew the solution. and was resolved much later, when it became clear that the infinite sum

$$\sum_{i=1}^{\infty} \frac{1}{2^i}$$

Converges to 1. Thus, this is now a falsidical paradox, something that surprises only those who still don't know about converging series.

This simple paradox becomes baffling when we compare it to Thomson's lamp paradox of Page 249. Assume that the runner has to cover a distance of one meter. In order to do that, he first has to cover 50 cm, but to do that he first has to cover 25 cm, and so on. On one hand, he can cover all this existence because their sum is one meter. On the other hand, at every point in the process he is short of his goal of one meter. How then can he be at one meter at the end. Also, how can there be an end to his infinite set of smaller and smaller distances?

This is similar to Thomson's lamp, which has no state at midnight. However, the runner is at the one meter point after traversing his infinite distances, because he must there! Obviously, he cannot be beyond that point, but why can't he be short of it? Being short of the one meter point means that he still has infinitely many short distances to run, but he has already run all the infinitely many distances.

The Greek philosopher Zeno of Elea (not to be confused with Zeno of Citium) is best known for his paradoxes which puzzled, challenged, influenced, inspired, infuriated, and amused his contemporaries as well as later generations. All have since been resolved.

The first Zeno paradox is referred to as the dichotomy paradox. It is related by Aristotle as follows "`That which is in locomotion must arrive at the half-way stage before it arrives at the goal.`" In other words, before we can cover a distance d, we first have to cover half that distance, but in order to move a distance of $d/2$, we first have to cover $d/4$, and so on. Thus, to cover any distance, we have to cover infinitely many distances, which seems impossible. Yet everyone knows that motion is possible, which results in a paradox.

To a contemporary reader, the resolution of this paradox is trivial. It is true that to move a distance d we have to cover infinitely many distances, but because each distance is half that of its predecessor, their sum converges.

In his 4th paradox, Zeno describes three rows of runners `X`, `Y`, and `Z`

$$\ldots \texttt{XXXXXXX} \ldots$$
$$\ldots \texttt{YYYYYYY} \ldots$$
$$\ldots \texttt{ZZZZZZZ} \ldots$$

The `XXX` runners are at rest, the `YYY` runners run at maximum speed to the left, and the `ZZZ` runners similarly run at maximum speed, but to the right. The paradox is that the *relative* speed between the `YYY` and `ZZZ` runners is *twice* the maximum speed.

This paradox was resolved in 1905 by the special theory of relativity, which states that there is a maximum speed c, the speed of light, but speeds should not be added arithmetically. Instead, special relativity provides the simple, unintuitive rule for adding speeds u and v

$$\frac{u+v}{1+\frac{u\,v}{c^2}}.$$

It is easy to see that when both u and v are maximum (both equal c, the combined speed is also c.

Another of Zeno paradoxes is concerned with repeatedly dividing a line segments into smaller and smaller segments. In the limit, the length of each segment is zero and there are infinitely many of them. To Zeno, the fact that a sum of zeros, even an infinite sum, produces a positive result, was a paradox. Today, we resolve this paradox by talking about the real numbers as non-denumerable, uncountable, or as forming a continuum.

- Galileo's paradox. Galileo, of the leaning tower of Pisa fame, noticed the mapping $n \leftrightarrow n^2$ that maps the positive integers to their squares. He realized that the set of squares of integers is a proper subset of the positive integers, so he concluded that this mapping constitutes a paradox, one that is named after him. Today, with the benefit of set theory, this paradox is resolved by simply saying that the sets of positive integers and of their squares have the same cardinality.
- A similar paradox of infinity is the Tristram Shandy paradox, resulting from story where Tristram Shandy decides to write his biography in such detail as to devote a day of writing to each day of his life. Contrary to any negative intuitive predictions, such a target is possible if he would live forever. See *The Life and Opinions of Tristram Shandy, Gentleman*, by Laurence Sterne, 1759.
- The preface paradox. Writing a book is not a trivial undertaking. The author normally checks his manuscript for errors before sending it to the publisher, where an editor, subeditor, associate editor, a production editor, and a copyeditor often also go

over the text, locate errors, make suggestions for improvements, and generally improve the text. After all this work, errors, mistakes, and mistypes may still be found in the book, especially if it is a technical textbook. Because of this, the author often includes a disclaimer In the preface to the book, admitting the inevitability of errors, apologizing for any errors readers may find, and promising to maintain a list of errors in the book's website.

If even one error is later found in the book, then the sentence or paragraph containing the error is wrong, and it causes the book to have a false statement. If no errors are ever found, then the inevitability of errors mentioned in the preface is itself a false statement. Either way, such a book contains a false statement. This is the preface paradox.

- The cable guy paradox. Your cable stopped working. The cable company has promised that a technician would come to your house tomorrow between 8AM and 4PM. This time window can be divided into the equal intervals $(8, 12]$ and $(12, 4)$, that we denote by AM and PM, respectively. If you had to bet on the technician's arrival, would you bet on AM or PM?

Clearly, at any time before 8AM, there is no reason to prefer any of the time intervals. After 4PM, the result would be known and there is no point in betting. Life experience teaches us that as you wait for the technician in the morning, you feel that the chance he would show up during AM is steadily shrinking, which is why during the AM interval you would tend to bet on PM. During the PM interval, it is pointless to bet AM.

Thus, in early morning, the AM and PM intervals have the same chance, and after that, you would tend to bet PM. The chance you would bet PM is therefore greater than 0.5, yet we know that half the time, a technician arrives in AM, and half the time he arrives in PM. This is the cable guy paradox.

- The barber paradox, a variation of Bertrand Russell's paradox, is an example of a falsidical paradox. This antinomy assumes that there is a town in which "the barber shaves all and only those men in town who do not shave themselves." Now we ask who shaves the barber? If he shaves himself, then he is shaving a man who shaves himself and violates the premise.
- Quinn's paradox. A poor person may sometimes be persuaded to donate blood plasma or to participate in an experimental medical procedure for money. Imagine a medical researcher who offers money for the following experiment. The subject is exposed to a mildly painful stimulus and is offered more money to agree to each additional step in the experiment, where the stimulus will be slightly increased. The steps are small enough such that the pain in each step is indistinguishable from the previous step.

On one hand, as the pain increases gradually, we expect the subject to stop the experiment at a certain point, feeling that the pain has become too much. On the other hand, if the steps are small enough, the subject may always accept the next step, because he knows from experience that the increased pain in the next step would be indistinguishable from the present pain. This is Quinn's paradox.

- Wang's paradox. This is a close relative of Quinn's paradox. It starts by stating "0 is small," continues with "if n is small then so is $n + 1$," and concludes by saying "so every number is small."

See also Exercise 6.2 for the sorites paradox.

- The rakehell paradox. The following quotation is from the novel *That Uncertain Feeling*, by Kingsley Amis (Penguin books, 1955).

"Feeling a tremendous rakehell, and not liking myself much for it, and feeling rather a good chap for not liking myself much for it, and not liking myself at all for feeling rather a good chap, I got indoors, vigorously rubbing lipstick off my mouth with my handkerchief."

(Rakehell means a dissolute man.) This is a paradox because the man in the story, being a rakehell, doesn't like himself, while also liking himself, because he has the guts to not like himself.

- Richard's paradox. We are familiar with words such as thousand, million, billion, and trillion, which describe large integers. In principle, it is possible to describe every integer in terms of such words. The description may specify how many trillions, millions, thousands, and so on, make up the integer. However, most real numbers are non-integers. They may have infinitely many digits after the decimal point, and it is not clear if it is possible to describe every real number in a finite number of words.

We therefore collect all the real numbers that can be described in a finite number of words in a set E. The following argument shows that E must be denumerable. Let n be the number of words used to describe an element R of E. We include R in a group with all the other members of E whose description consists of n words. The members of each group are arranged lexicographically. (This can be done even if a group is infinite and even if it is non-denumerable.) The groups are then opened and their members concatenated into one large list that contains all the members of E. Thus, set E is denumerable.

Assume that element R of E has the integer tag n. We can construct a new real number N whose integral part is zero and whose nth fractional digit is different from p, where p is the nth fractional digit of R. N is constructed in this way bit by bit from each R in E. Those who recall Cantor's diagonal argument from Page 255 will agree that N is different from any element of E. However, N has just been constructed by our finite description of it and must therefore belong to E; a contradiction.

This contradiction is Richard's paradox, first described by Jules Richard in 1905. See also Berry's Paradox, Page 342.

- König's paradox, by Julius König (1905), is closely related to Richard's paradox. We first notice that although the real numbers are non-denumerable (cannot be assigned tags), the entire set $\mathbb{R}$ of the reals is well-ordered and for any two reals it is possible to tell which is greater.

We again build a denumerable list E, a subset of $\mathbb{R}$ containing all the real numbers that can be described in a finite number of words. The elements of E are ordered as before, with tags assigned. Since $\mathbb{R}$ is well ordered, there must be a first real number (first, according to the order of the reals) which lies outside subset E and therefore cannot be defined by a finite number of words. But this number has just been defined by the previous sentences; a paradox.

- St Petersburg paradox. This involves simple probabilities. A player is willing to pay a fixed sum to play a simple game based on coin tossing. A coin is tossed again and again, landing on tails, until it lands on heads at the nth toss. The player then receives 2^n money units and the game is over.

This game and the resulting paradox were already discussed in 1738 by Daniel

Bernoulli. Reference [Feller 68] has a description and detailed statistical analysis of this game as well as the entire concept of what makes a game fair.

Intuition suggests that this is a game where the player has the advantage over the bank, because the chance of a fair coin to land heads is 0.5, and also because the longer it takes for it to land heads, the greater the gain for the player. We can therefore expect gamblers to be willing to pay handsomely, and to be happy to pay and play this game many times. However, because of human nature and of limited resources, there will always be a limit to what players are willing to risk in such a game. Here is a detailed analysis of the probabilities involved.

If a coin lands heads on the nth toss, it must have landed tails on the first $n-1$ tosses. The probability of this is 0.5^{n-1}. The probability of landing heads at step n is, of course, the same 0.5, so the chance of the coin landing heads on the nth toss but not before, is 0.5^n. This chance shrinks with the number of tosses, but since the payoff is 2^n, it grows, and it grows exponentially. Statistically speaking, we are dealing with independent random variables that grow exponentially as $2^1, 2^2, 2^3, \ldots$ with probabilities $2^{-1}, 2^{-2}, 2^{-3}, \ldots$.

Thus, the expectation of winning at step n is the product $2^n \times 2^{-n} = 1$ and the expectation of winning in n steps is the sum $\sum_1^n 1 = n$. This sum grows indefinitely, which reinforces our initial intuitive feeling that the player is the one who has the advantage in this game.

A variation of the game is to give the player a limit N such that he receives nothing if no heads is achieved after N steps. Now the gain is no longer infinite and is limited to the finite value N. This implies that the maximum a player should agree to pay to enter this game is N, but in practice there is also the time the game takes when N is large, and time, as we know, is money.

However, we know from experience with real players and gamblers, that there will always be a limit to what they are willing to pay for such a game, and this practical fact is the source of the St Petersburg paradox. Following are a few points that can be mentioned about this paradox.

1. In practice, this and similar games always have to be played with a limit, since the funds available to the bank are always limited. This resolves the paradox, because the gambler cannot expect infinite gains and would therefore not risk infinite money.

2. The concept of diminishing returns affects the amount a player is willing to risk. As the game is played repeatedly many times, the player gains more money. A time comes when large winnings, winnings that in the past were considered significant, are now considered puny. The player is rich and wants to get richer, but his gains are getting smaller and smaller the more money he has. A time may come when he decides to quit because the small gains may not be worth his time.

3. The definition of a gambler is someone who is attracted to risks, but most players in casinos are rational and try to avoid risks that they consider too high. Such a player avoids games where the gain is huge, but the chance of winning is small. Such a player may decide that the St Petersburg game is precisely of this type, and may decline to play.

4. Sophisticated gamblers, casino executives, statisticians, and others who analyze games and chances have developed the concept of a fair game. Here are a two informal definitions of what makes a game fair:

- A game is fair when the probability of winning is equal to the probability of losing.
- A game is fair if its expected value equals its cost.

Meaning of "fair game" in English: Someone or something that people are allowed to criticize. Celebrities are fair game for the press.

Thus, a fair game can be achieved by adjusting the rules, rewards, or conditions of the game so that the probabilities of each outcome are equal for all players. According to the definitions above, the St Petersburg game is unfair.

• Self-deception paradox. If I know what you are up to, you cannot deceive me. I always know what *I* am up to, so I can never deceive myself.

This simple statement proves by logic that a person cannot deceive himself, yet we know that people deceive themselves all the time. This is the self-deception paradox, whose most common example is self-denial. Psychologists and medical professionals have long realized that a terminal patient goes through several stages, one of them is self denial.

Fere libenter homines id quod volunt credunt (In general people willingly believe what they want to). —Julius Caesar, *Commentarii de Bello Gallico*, Book III, Ch. 18, 58–49 B.C.

◇ **Exercise 6.3:** What are other common examples of people deceiving themselves.

• Self-fulfilling belief paradox. Statement X states:

$$\text{I believe that I am believing statement } X.$$

Statement X is meaningless; it has no real content, yet if I believe in it, it becomes self-fulfilling. A placebo is a good example. A placebo is a harmless pill, medicine, or procedure prescribed more for the psychological benefit to the patient than for any physiological effect. Yet placebos often work if the patient believes that he has received something of value,

• The Unpleasant paradox. You are offered a large sum of money just to form in your mind the intention to swallow an unpleasant pill. Once you have formed this intention, you immediately receive the money, and you are still free to change your mind. This surprising feature is explained to you in advance.

The paradox is that you know in advance that you can change you mind and reverse your initial intention, yet not lose anything. However, this foreknowledge will prevent you from forming the intention in your mind in the first place. In other words, you cannot intend to do what you know you will not do.

• The Horror paradox. Script writers, movie directors, and television news producers have long discovered that horror movies, tragic news, and frightening stories hold a fascination for many readers and viewers. It is widely believed that the amount of pleasure and satisfaction they receive from such stories is proportional to the amount of unpleasantness and tragedy in the events themselves. The contrast between negative occurrences and positive feelings can be considered a paradox.

• The coin rotation paradox, the 1982 SAT question (almost) everyone got wrong. The top part of Figure 6.2 shows a circle rotating without slipping along a straight line

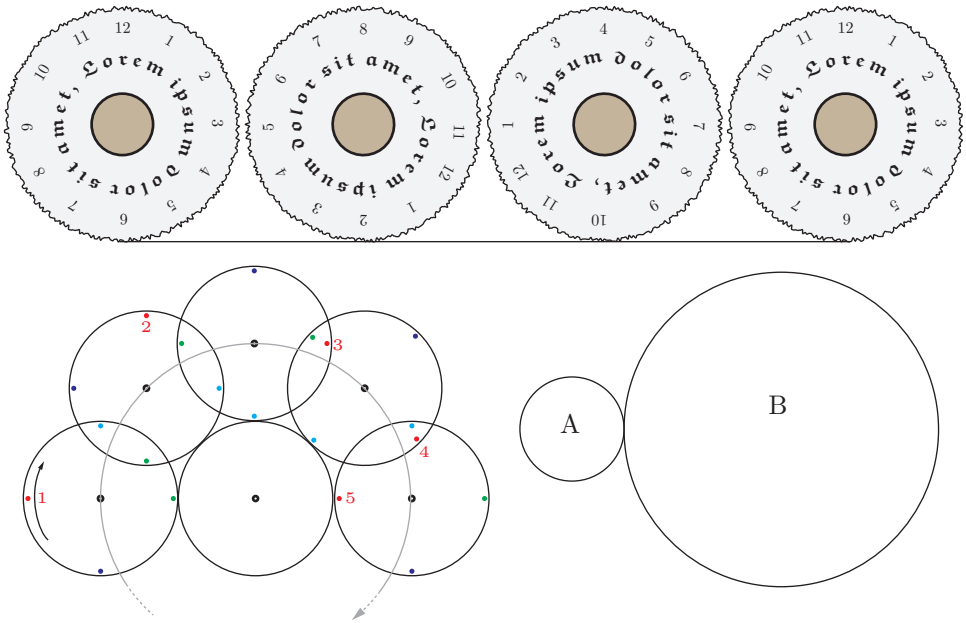


Figure 6.2: Rotating a Circle Along a Line and Around another Circle.

whose length equals the circumference of the circle. It is obvious that the circle makes a full, 360° turn.

The bottom-left part of the figure is an example of a compound rotation. It shows a circle rotating around the circumference of another, central circle of the same radius. Five positions are shown and they illustrate the surprising, unintuitive fact that after a full rotation, the rotated circle has made *two* full rotations, not one. This is the result of its rotating about its center while also making a full revolution about the center of the central circle.

It is easy to extend this problem to circles of different radii. The bottom-right part of the figure shows a central circle B whose radius (and therefore also its circumference) is three times greater than that of the rotating circle A around it. Rotating A along the straight circumference of B, it would make three rotations, but the extra rotation about the center of B results in a total of four rotations. In general, the number of rotations is one more than the ratio of the two radii.

- Achilles and the fly paradox. According to Martin Gardner, this interesting paradox originated by A. K. Austin of Sheffield, who published it in *Mathematics Magazine* in 1971. In one of the famous Zeno paradoxes, Achilles is pursuing the tortoise and will never overtake it. (See also the racecourse paradox on Page 319.) In this modern version, Achilles, running at 5 mph, is pursuing the tortoise, which is slow-moving at 1 mph, but had a head start. At the point in time when Achilles catches up with the tortoise, a fly, flying at 10 mph, arrives at the meeting point and starts flying back and forth between Achilles and the tortoise as the distance between them grows. After one hour, the distance between Achilles and the tortoise is 4 miles, and the question is, where is



Figure 6.3: Image by MS Copilot, Sept. 2025.

the fly?

The unexpected answer, which is also the source of the paradox, is that the fly can be anywhere between them. This is explained in two steps as follows:

Step 1. When the three start their shared running/flying trip at the meeting point, the distance between them is zero. As they move in time, the distance is growing, but initially it is very small. The fly must therefore reverse its direction right from the point of meeting, and in fact it has to reverse its direction infinitely many times at any time interval since the beginning.

To make this picture more rigorous, imagine the situation after a small interval of ϵ time units. Achilles has moved a distance of 5ϵ, the tortoise has moved one ϵ, and the distance between them is the small 4ϵ. In order to stay between the two runners, the fly, which has flown 10ϵ, must already have reversed its direction infinitely many times. Since the ϵ time interval is arbitrary and may be as short as we wish, it follows that the fly must reverse its direction infinitely many times right from the meeting point. There is no initial non-zero interval during which it flies straight without reversing his direction. This situation is similar to the Thomson's lamp paradox, Page 249.

Step 2. Going back in time from an hour after the meeting to the point of the meeting. Assume that an hour after the meeting the fly is at an arbitrary point between the two protagonists. As we move back in time, the fly must reverse directions more and more often and fly shorter and shorter segments. Right before it approaches the meeting point, the fly must reverse its direction infinitely many times, which implies that it has no first direction. Hence, it will traverse back in time and reach the meeting point in space as time reaches the meeting time, regardless of where it was at the end of the hour (i.e., when the time reversal started). The inescapable conclusion is that at time ϵ, or at one second, or at the end of an hour, or in fact at any time—the fly could be anywhere between Achilles and the tortoise.

There is another unusual feature to this hypothetical race. At the time of meeting, the fly is already flying. It has reached the meeting point from a certain direction. Hence, at that time it has a well-defined velocity (a vector describing its speed and direction). During the hour-long race, the fly's speed is constant, but its direction

reverses all the time. Its velocity is therefore a discontinuous function that has sharp changes at infinitely many points where it is also non-differentiable. Acceleration is the derivative of velocity, so when the velocity vector changes from the constant 10 mph to the constant -10 mph, the acceleration vector must change from zero to $-\infty$ (stop) and then instantaneously to $+\infty$ (start). This acceleration function is discontinuous at infinitely many points.

- An example of a veridical paradox is Gabriel's horn (also called Torricelli's trumpet). This is a three-dimensional geometric figure that has infinite surface area but finite volume. Its name originates in the traditional belief that the archangel Gabriel will blow his horn to announce Judgment Day. Evangelista Torricelli seems to have been the first to rigorously study, in the 17th century, the properties of this object.

The horn is the figure obtained when the curve $y = 1/x$ is drawn from $x = 1$ to $x = a$ for some positive value $a > 1$, and is then rotated 360° about the x-axis. Reference [WikiHorn 23] proves that the volume of the horn is $\pi(1 - 1/a)$, while its surface area is $2\pi \ln a$. Thus, for a at infinity, the volume is the very finite π, while the surface area is infinite. In simple terms, this horn is infinitely long but has a finite volume. It seems to connect the finite with the infinite, or at least to blur the boundary between them; unexpected, weird, and surprising. This is a veridical paradox because it is a surprise that tends to evaporate as we ponder it and get used to its unusual but true results. Here are a few more words that may help the reader to come to terms with this surprising horn.

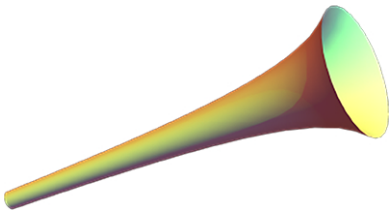


Figure 6.4: Gabriel's Horn.

```
a=10; (*Gabriel's Horn*)
ParametricPlot3D[{2Cos[u]/w, ( 2Sin[u])/w, w}, {u,0,2Pi},
 {w,1.5,a}, PlotStyle->{LightOrange,Specularity[Green,2]},
 Boxed->False,Axes->None,Mesh->None]
```

A simple explanation of the finite volume of the horn is to consider it a series of infinitely many thin disks of diminishing radii. The radius of a disk located at coordinate x is $r = 1/x$, so its area is $\pi r^2 = \pi/x^2$. The volume of the horn is the infinite sum of the disk areas and this sum converges because the infinite sum of $1/x^2$ converges. (The exact volume is obtained when the disks are infinitely thin, which requires a definite integral instead of an infinite sum.)

> An infinite number of mathematicians walk into a bar. The first one orders a pint. The second one orders half a pint. The next ones order a quarter, then an eighth, and so on. The bartender hands them two pints and says "You mathematicians, you don't know your limits."
>
> —A mathematical joke.

The fact that the horn has finite volume but infinite surface area gives rise to the painter's paradox, another veridical paradox. It takes a small, finite amount of paint to fill the horn, and this paint already covers its interior surface. Yet it takes an infinite amount of paint to actually paint the infinite (interior or exterior) surface. This paradox is veridical because it disappears when we understand that paint in the real world has finite thickness, but the paint needed for the horn must be infinitely thin. Trying to fill the horn with such paint would also require an infinite amount.

Another way to explain this paradox is to understand the meaning of the term dimension. When we say that the length of a line is a foot, we have simply compared the line to a known object, a ruler or a tape measure, marked in feet. Without a standard of length such as a ruler, it is impossible to measure any length and assign it a value. Also, when comparing dimensions, it is meaningful only to compare dimensions of the same type. It is meaningless to compare apples and oranges, to compare time to mass, or to compare surface to volume.

We intuitively feel that the volume of an object is greater than its length or surface area, but this is only a feeling we got from our experience with everyday objects; it is not a precise concept. Therefore, volume and area cannot be compared. A similar example of dimensions that are not commensurate is the fractal known as the Koch snowflake. It has a finite area, but its boundary curve is so fragmented, jagged, crinkled, and zigzaggy that its length is infinite.

The guiding principle in cases where we are surprised by seemingly impossible mathematical results is to remember that mathematics is a creation of the human mind and is not restricted by the features and limitations of the real world.

◇ **Exercise 6.4:** Come up with an everyday object that is easy to construct, but when carried to infinity it features the same infinite area and finite volume.

• Another example of a veridical paradox is a someone who is 20 years old but has had only five birthdays. The solution is that his birthday is on a Leap Day, February 29th, and only occurs once every four years. Such a case is logically consistent, but it could be argued that the statement is nonsensical.

◇ **Exercise 6.5:** Two baby girls were born on the same day, to the same mother, in the same hospital, yet they are not twins. How can this be?

• Potato Paradoxes. The potato paradox, sometimes called the potato effect or potato puzzle, is a group of similar, veridical paradoxes described in reference [WikiPotato 25]. A fresh potato consists of 90% water. You leave it until its water content drops to 98%. How much does it weigh now? The surprising answer is half its original weight.

The version described here starts with 100 persons in a room, 99 of whom are left-handed. How many left-handed persons must leave the room in order for this percentage to drop to 98? The intuitive answer is 1, but if one person leaves, the ratio of left-handed

The Room of Raised Hands

The hall was vast, echoing with quiet anticipation. A hundred souls filled the space—young children with wide eyes, elders with weathered faces, and every age in between. Their skin tones ranged from the pale frost of the Arctic to the deep warmth of the Sahara. Their garments told stories: shimmering Indian saris, crisp military uniforms, flowing Muslim robes, swimsuits still damp from the sea, and sacred Jewish talits draped with reverence.

They had come from every corner of the Earth. Not for war. Not for politics. But for something older than both, understanding.

At the front stood a young girl. Her name was not important, nor her nationality. What mattered was her silence. She looked out at the crowd, and slowly, deliberately, raised her right hand.

A hush fell.

Then, one by one, the others raised their left hands. Not in defiance, not in mimicry, but in harmony. It was as if the room had rehearsed this moment for centuries.

The gesture was simple. Yet it carried weight.

The girl's right hand symbolized the individual, the lone voice, the unique perspective. The crowd's left hands symbolized the collective, the shared hope, the common dream.

And in that moment, the room was not divided. It was balanced.

A child whispered to her grandmother, "Why did she raise the other hand?"

The grandmother smiled. "Because sometimes, to be heard, you must be different. And sometimes, to listen, you must be many."



Figure 6.5: Image and Text by MS Copilot, Aug. 2025.

persons becomes $98/99 \approx 98.9898$. Surprisingly, the ratio has only dropped a bit from 99, and is still quite a ways from 98. Even when 10 people leave the room, the ratio would still be $89/90 \approx 98.8888$. It is only after 50 people have left, that the ratio is finally down to $49/50 = 0.98$. Why is this so unintuitive?

A logical approach to the problem starts with the realization that the number of right-handed persons in the room is 1, and it will not change when more and more left-handed persons leave. Thus, for the ratio to be 0.98, which equals $1 - 0.02$, this right-handed person must consist of 0.02 or 2% of the entire room, and 2% of 100 is 50.

A formal way of solving this is, of course, to set an equation, and this happens to be simple. The ratio of left-handed people after x people have left the room is $(99 - x)/(100 - x)$, and when we equate this ratio to 0.98, it solves to $x = 1/0.02 = 50$.

• A self-contradictory phrase such as "There is no absolute truth" can be considered an antinomy because this statement seems absolutely true, and therefore denies itself. Other antimonies are: You can't lose what you never had. "You must unlearn, what you have learned" (Yoda). The restaurant owner said "the customer is always right and I am wrong" and the customer answered "you're right." If an all powerful being, an omnipotence, creates an immovable object, then by definition, he cannot move it. He has created an object that even he is not powerful enough to move, thus ceasing to be omnipotent. The liar paradox: Pinocchio says: "My nose will grow." If his nose is growing, then he wasn't lying, so his nose shouldn't have grown. If his nose doesn't grow, then he was lying, but in this case his nose should have grown. (A simpler version of this paradox is: "this sentence is false.") Time travel is a fertile source of paradoxes. Perhaps the most obvious is for a person to go back in time and kill his grandfather before his father is born. I'll leave you to ponder the implications of this.

◇ **Exercise 6.6:** What is the original version of this paradox?

The liar paradox is well known because it has many relatives. Here are a few:

▪ Consider the infinite sequence of identical statements "All the following statements are untrue." It turns out that it is impossible to assign consistent `true` or `false` values to these statements.

▪ The following two statements are cyclic and therefore self referential:
Socrates: "What Plato is saying is false."
Plato: "What Socrates is saying is true."

▪ The following finite sequence of statements alternate between true and false. If the last statement is true, then the first one must be untrue and vice versa. The precise values depends on whether n is odd or even.

1 The following sentence is untrue.
2 The following sentence is untrue.
3 The following sentence is untrue.
⋮
n The first sentence is true.

• Aristotle's wheel paradox This paradox has to do with physical objects, namely rolling wheels. A general reference is [wikiAristoPardox 23]. Consider the wheels of

Figure 6.6. As shown in part (c) of the figure, such a wheel consists of two parts, large (R), and small (r), that are attached and must spin together like the wheel of a train (imagine a dime glued to a half-dollar). Assume that the wheel on the left of part (a) is rolled without slipping (or skidding) one full revolution to the right. The figure implies that both points 1 and 2 travel horizontally a distance of $2\pi R$ units (the circumference of the large wheel), but a little thinking suggests that point 2 should travel only the circumference $2\pi r$ of the smaller part (in red); a paradox!

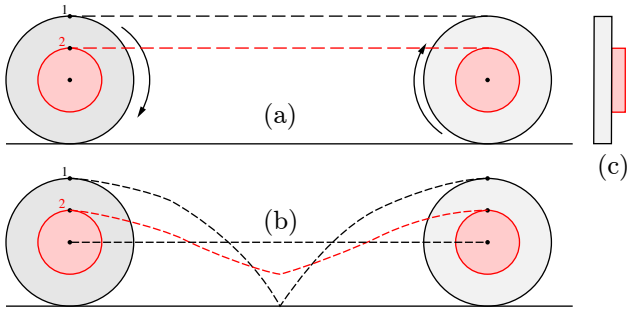


Figure 6.6: The Aristotle Paradox.

The same paradox appears if the small wheel rolls on a narrow track and causes the large wheel to roll with it, as happens, for example, in trains wheels. In such a wheel, the small part rides on the rail while the large part is a flange which protrudes down the side of the rail.

This paradox is attributed to Aristotle. For centuries, it has baffled mathematicians and philosophers such as Cardano, Mersenne, Galileo, and Bolzano, but to our amazement, the solution is quite simple and is presented here in both a physical and mathematical form.

- Physical solution. The wheel must be rolled on the track without slipping or skidding, but this is true only for point 1, which is on the circumference of the large wheel. Point 2, on the other hand, is both rotating and slipping all the time. It is slipping with respect to the track. This slipping is not noticed by the viewer of a diagram or the observer of a video or real wheel rolling, because the rim of the small wheel does not touch anything. However, the following simple experiment can reveal the slipping. Roll the large wheel along its track while holding another, narrow track, such as a rough strip of wood or plastic, to the rim of the small wheel.
- Mathematical solution. Part (b) of Figure 6.6 illustrates the subtle fact that when the wheel is rolled, both points 1 and 2 move first down and later up in addition to moving horizontally. It has been known for centuries that a point attached to the rim of a rolling wheel moves along a curve called a cycloid. It is obvious from the figure that the cycloid path of point 2 is shorter than that of point 1, and that the path plotted by the center of the wheel is the straight-line limit of the black, outermost cycloid. Thus, all points between the rim and the center of the wheel move along cycloids whose lengths get shorter as we move toward the center, and there is no paradox.

- A game with a simple logical paradox. Two envelopes contain money, one twice as much as the other. We pick up one envelope to win its content, but before we open it, we are given the option of swapping the envelopes. This seems pointless, because the only information available is the amounts of money, say, M and $2M$, but consider the following. If the original envelope we chose contain X amount, then the other envelope contain either $2X$ or $X/2$. The expected value (EV) of that envelope is therefore $\text{EV} = (2X + X/2)/2 = 1.5X$, which is greater than X! This is a paradox, because once we swap envelopes, we would be facing the same dilemma!

  The solution is that there is really no advantage in swapping. We only need to realize that the two X values in the above expression do not refer to the same amount. In the case where swapping is going to double our money, the X (what is in the original envelope) is the smaller of the two amounts, while if by swapping we are going to lose, the X is the larger amount. Suppose the amounts are 100 and 200. The average, which is also the expected value EV, is 150. When we swap, the $2X$ amount is 100 while the $X/2$ amount is 200. Thus, the EV in this case is $[2 \cdot 100 + \frac{200}{2}]/2 = 150$, the same as the EV with the original envelope.
- The friendship paradox. In a set of people that are not complete strangers, there may be relations that we can generally group under the title "friendship." A member of the set may have friends, neighbors, partners, relatives, or other persons toward which he has special feelings. The paradox discussed here is due to Scott L. Feld [Feld 91]. From his surveys of social interactions of real people he concluded that on average, the friends of a set member A have more friends than A himself has.

  This is unexpected and it seems to contradict common sense. What is special about A's friends and why should they be friendlier than A himself? After all, A was chosen at random in the set. One explanation is that friendly people, people with many friends, are simply more likely to also be A's friends. Similarly, if B has few friends, it is unlikely that A would be one of them. A more rigorous explanation is given on page 1470 of [Feld 91], where Feld derives a simple expression that proves his claim.

L'ami de mon amie (the friend of my friend) is a 1987 movie by the French director Éric Rohmer.

- The Dartboard Paradox.



Dartboards are popular entertainment in certain parts of the world. The following information is from `darting.com`. The standard international dartboard is known as the "clock" board. The standard dartboard measurement is 18 inches in diameter and has 20 numbered, pie-shaped segments of equal size, plus the bull's-eye. Each numbered segment has a double ring on the outside perimeter of the scoring area and an interior triple ring (which count as double and triple of that number respectively). The bull's-eye has an outerbull area (also known as the single bull, which scores 25) and an inner bull (also known as a double bull's-eye, which scores 50).

But our interest in darting is mathematical and it involves probabilities. We consider the dartboard a mathematical circle and the dart tip as a mathematical point. Since the board is circular, it makes sense to use polar coordinates (r, θ) and assign two

real numbers to each point on the board. The number of points is not just infinite, but uncountable, which is why the probability of the dart hitting any point at random is zero, yet an experienced player can hit a point on the board most of the time.

This behavior of the board (as well as any uncountable space, a continuum) is baffling. We normally work with discrete probability spaces, such as coin toss, die roll, or choosing an integer between 0 and 100 at random. In such a space, each result has a probability, a real number in the interval $[0, 1]$, and the probabilities add up to 1.

In an infinite but countable space of events, there is an infinite number of probabilities, but they may also make sense because we know that certain infinite series converge to a finite value. However, in a continuum, an infinite series of probabilities, i.e., nonnegative values, grows without bounds, has an infinite sum, and cannot converge. (Except if it has only a finite number of nonzero values or a countable number of positive values that shrink and converge.)

Thus, probabilities in an uncountable space, a continuum, present a paradox that needs to be resolved. The following discussion leads to the concept of probability density, the probability of a group or a range of events, which is considered a satisfactory resolution of this paradox.

We start with familiar concepts. How to compute probabilities in a discrete space, where there is either a finite number or an infinite but countable number of cases. Examples are a coin toss (two results), rolling a die (six cases), drawing a card (one of 52 possibilities), or choosing a number in a large lottery (millions of cases). In such cases, the probability that a value chosen at random will fall into a given set of cases is the sum of the individual probabilities.

In drawing a card, for example, the probability $P(x \in \{\clubsuit, \heartsuit, \spadesuit\})$ of a draw producing a clubsuit, or a heartsuit, or a spadesuit is the sum of the individual probabilities $P(x = \clubsuit) + P(x = \heartsuit) + P(x = \spadesuit) = 3/4$. (As always, we have to be careful when adding probabilities.)

In the case of an infinite but countable space of probabilities, we can still apply the same rule, but only if the individual probabilities converge to a finite sum. If the sum converges, we can normalize the individual probabilities such that their sum would be 1. Consider a process that proceeds in steps, where the probability of some outcome is halved in each step. If the probabilities of the first few steps are p, $p/2$, $p/4$,..., we can normalize them by dividing each by $2p$ to obtain the sequence of normalized probabilities $1/2$, $1/4$, $1/8$, ..., $1/2^k$, ..., whose sum is 1. In such a case, we can add probabilities to find the probability of a set of events. The probability of the set of events 1, 2, and 8 would then be the sum

$$1/2 + 1/2^2 + 1/2^8 = \frac{2^7 + 2^6 + 2^0}{2^8} = 193/256 = 0.7539.$$

◇ **Exercise 6.7:** What could be an example of such a process?

However, these arguments break down when dealing with a continuum, where there is an uncountable number of cases or results. Such a case presents a different situation. A set of probabilities that is halved in every step cannot exist here, because in a continuum it is impossible to proceed in individual steps. A step in a continuum does not have an immediate successor or predecessor, much as a real number does not have them. Stated

differently, there cannot be an infinite convergent series of probabilities of the form $1/2^k$ (or any other form) simply because in a continuum of events, an event cannot have an integer tag k. There are infinitely many integers, yet there are "not enough" integers to tag every real number or every element of a continuum.

Thus, when dealing with a continuum, we seem to have a paradox. We can no longer talk about the probability of an individual case or result. Given any interval of real numbers, we can easily choose any number R in it at random, even though the probability of choosing it is zero! This probability cannot be positive, because that would cause the total probability of choosing *a* number, any number, in the interval—0.4 or 0.67 or 0.88 or 0.601 or any other—to be infinite, and probabilities must be in the interval $[0, 1]$. The paradox is that when we choose a number R at random, we have chosen it even though its probability to be chosen is zero!

Therefore, in a continuum, we can only consider the probability of a range of cases. This important idea resolves our paradox by taking us from the basic concept of probability to the advanced concept of probability density (PDF), the probability of an interval of values or events rather than that of an individual value.

Given an interval such as $[0, 1]$, the probability of choosing any one of its elements is 1, because we know that we can choose 0.5 or 0.46, or any other element. Also, the probability of choosing a number in a subinterval such as $[0.55, 0.65]$ is 0.1, because this is the ratio of the length (0.1) of the subinterval to the length (1) of the entire interval. We can therefore think of the probability of the subinterval $[0.55, 0.65]$ as a rectangle of height 1, centered on the subinterval, whose width is $0.65 - 0.55 = 0.1$. Its area is therefore 0.1 and it equals the probability of choosing a real number at random out of this subinterval.

So far we have assumed an interval with uniform probability; every item has the same probability. In practice, there may be a probability curve that determines the probability of an event. Going back to the dartboard, assume that someone devious has hidden a magnet behind the board (Figure 6.7b), which attracts the darts and pulls them closer to a point away from the bull's-eye. For simplicity, let's consider the one-dimensional interval $[0, 1]$, rather than the two-dimensional circular dartboard, and assume that events, marked by real numbers in this interval, have probabilities as indicated by Figure 6.7a,b. These probabilities, which are defined by the function $x^3(1 - x^3)$, start at zero, climb and reach a maximum at $x = 0.793701$, and then drop smoothly down to zero at $x = 1$.

This is an example of nonuniform probability. The probabilities of subintervals are different, so we try to look for probability density rather than a probability at a point and we concentrate on the subinterval $[0.55, 0.65]$ marked in the figure in cyan. The Mathematica code listed with the figure indicates that the curve reaches its maximum at $x = 0.793701$. The two integrals that follow show that the two areas under the curve (1) over the entire interval and (2) over the subinterval are 0.107143 and 0.0169217, respectively. In order to normalize the area 0.107143 to 1, it has to be multiplied by a factor of (approximately) 9.333. When the area 0.0169217 is multiplied by the same factor, the result, 0.157936, is the probability of choosing a point at random on the curve in the subinterval $[0.55, 0.65]$ (or, alternatively, the probability density of any of the events that correspond to this subinterval).

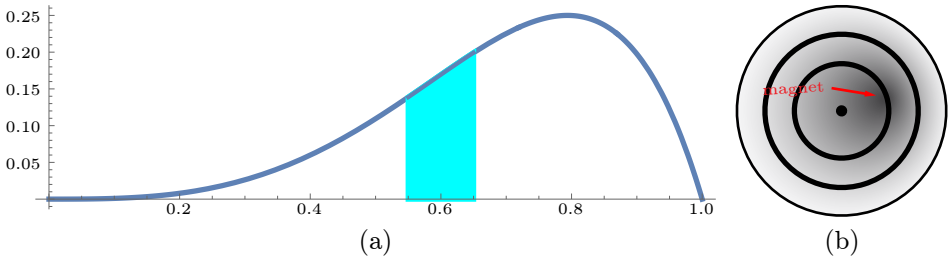


Figure 6.7: Probability of a subinterval.

```
(*Probability of a subinterval in a continuous interval*)
crv[x_]:=x^3(1.-x^3);g1=Plot[crv[x], {x,0,1},
 PlotStyle->{Thick}, AspectRatio->.3];
g2=Plot[crv[x], {x,0.55,0.65}, AspectRatio->.3,
 Filling->0,FillingStyle->{Cyan}];
Show[g1,g2]
Solve[D[crv[x], x]==0] (* Find point of maximum *)
a = Integrate[crv[x], {x, 0, 1.}]
b = Integrate[crv[x], {x, 0.55, 0.65}]
b/a
```

◇ **Exercise 6.8:** Compute the probability of choosing a point at random on the same curve in the subinterval $[0.75, 0.85]$. Why is the result different?

We end with a quotation from Wikipedia, reference [WikiDensty 23]: "In a more precise sense, the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variable's PDF over that range, that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and the area under the entire curve is equal to 1."

The Sierpinski-Mazurkiewicz paradox, found on Page 249, is another surprising result that stems from the nature of the infinite.

- The existence of non-collision singularities. In Euclid's geometry, a point is that which has no dimensions. When Newton developed the science of mechanics, he did not consider the finite size of atoms and elementary particles, and the fact that dimensionless particles can result in infinities in physical calculations. Newton wrote the law of universal gravitation in the familiar form

$$F = G\frac{mM}{r^2}, \tag{6.1}$$

ignoring the fact that if r, the distance between two material objects, can shrink down to mathematical zero, the gravitational force between the objects becomes infinite. (His unquestioned genius notwithstanding, Newton was a man of his time, and often ignored rigor in his mathematics.)

In general relativity, on the other hand, speed is limited to the speed of light and distances are limited to $c^4/4G$, thereby removing the paradox of zero distance and infinite force.

The 1973 paradox described here, due to Jeff Xia, is titled "The existence of non-collision singularities in a Newtonian system" [Xia 92]. It easy to understand and is surprising, but its resolution is obvious once we realize the differences between Newtonian gravitation and general relativity. This paradox is described here in three steps, considering gravitational systems of two, three, and five particles.

1. A two-particle system is easy to compute and to solve using Newton's methods, and such calculations are routinely done in order to predict the positions of a planet around a star or a moon orbiting a planet. A system of three particles (the so-called three-body problem) presents a completely different problem. There are closed expressions for solving such a problem but they involve infinite sums, which is why such a problem is handled by numerically calculating the velocity and position of each of the three objects in small steps; a simulation in time.

2. However, there is a well-known special case of the three-body problem. It starts with a pair of objects, a binary, of equal masses m_1 and m_2, each orbiting their common center of gravity (an empty point g in space). We call their plane of orbit the xy plane and assume that there is a third object, of a much smaller mass m_3, moving along the z axis toward point g as in Figure 6.8a, gravitationally attracted by m_1 and m_2. As long as the two objects move in small circles, their distances from m_3 are equal, so both pull it with the same force, a force that depends on the distance of m_3 from the xy plane.

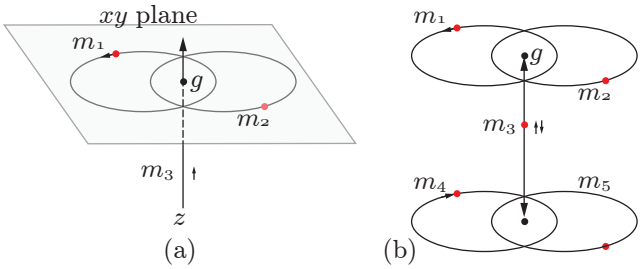


Figure 6.8: A Five-Body Problem.

If, however, the two objects orbit in ellipses, especially ellipses with large eccentricities, then the forces they apply on m_3 depend not just on its distance from the xy plane but also on the distance between m_1 and m_2, a distance that varies widely and continuously as they orbit point g. This is especially noticeable in the extreme case where their elliptic orbits resemble a straight line and they can be either very far from each other or arbitrarily close.

In such a case, m_3 is attracted to the xy plane, it is accelerated toward it, and eventually passes through point g and continues above the plane. If at that point m_1 and m_2 are very close, they start pulling m_3 down with a strong force (a force that can be made as strong as desired by adjusting their masses and the distances between them). Thus, m_3 stops and starts going down fast. While it does that, the binary separates

and moves away in their orbits, thereby reducing their pull on m_3, which then continues down the z axis and may, given the right conditions, be completely ejected from the binary system and escape. An important side effect is that reversing the direction of m_3 causes the binary pair to move upward a short distance; to recoil, an effect that is demanded by the law of conservation of momentum.

3. The paradox appears in references [Xia 92] and [Saari 95], where m_3 is prevented from escaping by adding more masses to the original gravitational system. Adding a fourth mass, it would have to be placed on the z-axis and would simply pull m_3 toward it, ending with a collision. At least two more masses m_4 and m_5, are therefore needed. They move in orbits similar to those of m_1 and m_2, on a plane parallel to the xy plane but a distance below it as in Figure 6.8b.

By adjusting the masses, distances, and orbits (a delicate process, like balancing a needle on its point), it is possible to end up with a five-body system where m_3 oscillates up and down along the z-axis, passing the xy plane, stopping, going down to the m_4m_5 plane (while causing m_1 and m_2 to go up slightly), passing it, stopping, and going up while causing m_4 and m_5 to move down slightly. It then bounces between the planes of the binaries, causing them to separate further away each time.

The surprise is that after an infinite number of oscillations, the two binary pairs are pushed to an infinite distance from each other, but all this (the infinite oscillations and recoils) happens *in a finite amount of time!*

The resolution of this paradox is obvious when we realize that Newton's theory of gravitation, even though very useful and basically correct, has something wrong deep within it. It implicitly assumes that particles are dimensionless; they are mathematical points and can therefore touch each other, causing the gravitational force between them, expressed by Equation (6.1), to become infinite. In contrast, the 1916 theory of general relativity posits that c, the speed of light in a vacuum, is the limit of speed and that there is a maximum force that is given by $c^4/4G$. These limitations guarantee that the particles in the five-body system above cannot go to infinity in a finite amount of time. Instead, when you try to bring two masses too close together, a black hole horizon forms around them; they become a black hole.

- Bertrand paradox. The Bertrand paradox, introduced by Joseph Bertrand in 1889, has to do with the computation of probabilities, where different methods of choosing the relevant objects lead to different results. The problem is easy to describe and understand. See, for example, reference [Bertrand paradox 24]. An equilateral triangle is inscribed in a circle. A chord is chosen at random in the circle. What is the probability that the chord is longer than a side of the triangle? Bertrand himself provided three solutions, all seemingly correct, and all leading to different probability values. Here they are.

Method 1. Random endpoints. Two points are chosen at random on the circumference of the circle. They define a unique chord. The triangle is rotated such that one of its vertices coincides with one of the two points, as in Figure 6.9a. No generality is lost in this process, but it is now easy to make a decision just by examining the other endpoint of the chord. If this point lies on the arc between the endpoints of the triangle side opposite the first point, then the chord (red in the figure) is longer than a side of the triangle. The length of that arc (green in the figure) is one third of the circumference of the circle, implying that the probability in question is 1/3.

Method 2. Random radial point. A line segment is down from the center C of the

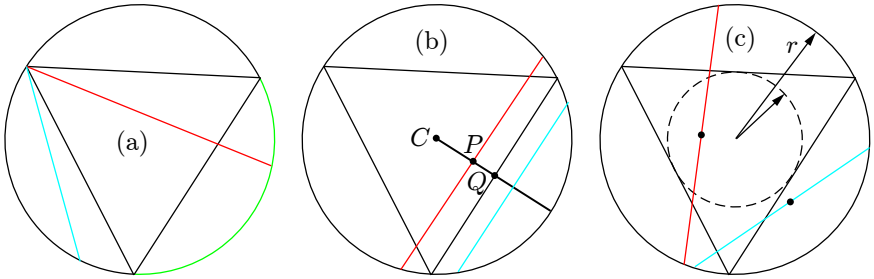


Figure 6.9: Probabilities in Bertrand Paradox.

circle to the circumference. It becomes a radius, as shown in part (b) of the figure. A point P is chosen any random on the radius to become the midpoint of a chord which is drawn perpendicular to the radius. The circle is rotated, with the triangle, so that one side of the triangle is perpendicular to the radius. The point where this side intersects the radius is denoted by Q. If $CP < CQ$, then the chord (red in the figure) is longer than the triangle side. However, because the triangle is equilateral, Q is also the midpoint of the radius, which implies that half the time CP will be less than CQ and half the time CP will be greater than CQ. The probability derived by this method is therefore $1/2$.

Method 3. Random midpoint. A point P is randomly chosen within the circle whose radius is denoted by r. A chord is drawn with P as its midpoint. The chord will be longer than a side of the triangle if P happens to be located within a concentric circle of radius $r/2$. It is now clear that this method yields a probability of $1/4$, because the area of the inner circle is $1/4$ the area of the original circle.

The same problem seems to have three different solutions, which is a paradox. The first attempt at a resolution is often to use our powerful, fast computers to simulate the problem by drawing random (actually, pseudorandom) numbers, and produce an approximate numeric result. However, when the three methods above are programmed and run many times, each produces the theoretical value obtained earlier. The paradox still exists.

Careful scrutiny and analysis explain that the paradox is the result of counting the diameters (the longest chords). The different geometric constructions used by the three methods result in counting too many or too few diameters, which affect the computed probabilities. Here is why.

Method 2 shows preference toward long chords, because it chooses each diameter in two ways, while choosing any other chord uniquely.

Method 3 is based on the observation that a chord is uniquely determined by the location of its midpoint in the circle.

This method results in the lower probability of $1/4$ because it misses diameters. A diameter is a chord whose midpoint is the center of the circle. The center, however, is the midpoint of all the infinitely many diameters, but choosing the center at random is rare, which is why this method misses diameters. Any other random point inside the circle corresponds to exactly one chord.

Bertrand's resolution of this paradox says that special care must be taken when considering probabilities in an infinite space. It may happen that a statement such as

"choose a random ..." may not be well defined and it is possible to choose the ... in different ways that eventually result in different probabilities. Thus, Bertrand concludes that the method of selecting chords at random should be specified as part the problem, not as part of the solution. Instead of having a problem with three different solutions (a paradox), there really are three problems, each with its own solution.

Another approach to a resolution is to claim that there is a unique solution to the problem and we can find it in the following process.

1. Fill up the infinite two-dimensional space with straight lines.

2. Draw a circle of radius r somewhere and construct an inscribed equilateral triangle t.

3. Some of the lines created in step 1 pass through the circle and define chords. Count the number of chords that are longer than a side of t and divide by the total number of chords in the circle. If those numbers is infinite, approach the quotient as a limit.

4. Now move the circle and perhaps also vary its radius many times, each time performing the computations of step 3. If the results are consistent, it proves that the desired probability exists and is well defined.

Notice that this process implicitly assumes that a chord is an infinite line that passes through a circle and not a finite segment that touches two points on the circumference.

◇ **Exercise 6.9:** Here is a seemingly foolproof way of generating a random chord in a circle. Draw all the possible chords in the circle. Assign each chord a number, and choose one number at random. No geometric construction is needed, and no diameters are omitted or counted multiple times. What is wrong with this idea?

Yet another approach to a resolution is more philosophical. One could say: We use probability in cases where there is not enough information to obtain a precise result. Does this always permit us to use precise numbers between 0 and 1 for our probabilities? Can we always trust mathematics and let it assign cold numbers to our ignorance? In Bertrand's case, perhaps the best that we can do is take an average of the three solutions above and claims that $(1/2 + 1/3 + 1/4)/3 = 13/36$ is the best result that we can hope for.

• Bertrand's box paradox, not to be confused with Bertrand paradox, is somewhat similar to the Monty Hall problem, (Page 341). Introduced by Joseph Bertrand, also in 1889, it is easy to understand and solve once we realize that this is an example of conditional probability, so here is a short discussion of this important topic.

Conditional probability is the case where the probability of an event B is affected by conditions A that are related to B and affect it. Bayes' theorem (or Bayes' rule), named after Thomas Bayes, describes the conditional probability $P(A|B)$, the probability of event A occurring given that B is true, by the simple expression

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where (1) A and B are events and $P(B) \neq 0$, (2) $P(A|B)$ is a conditional probability; the probability of event A occurring given that B is true, (3) $P(B|A)$ is the probability of event B occurring given that A is true (it can also be interpreted as the likelihood

of A given a fixed B), and (4) $P(A)$ and $P(B)$ are the probabilities of events A and B respectively without any affecting conditions.

■ **Examples**. A game—such as sport, chess, or cards—played by two teams. It is obvious that the presence of a major player in one of the teams is likely to affect the outcome.

■ This is actually a large set of examples. A set of objects is given. What is the probability that all the objects of the set have some property, given that at least one of them does.

■ A die is rolled and you are told that the result was greater than four. What is the probability of the result being an even number?

■ A bag has known numbers of red, green, and blue marbles. A marble has just been drawn and put back, and you are told that it was not green. What is the probability that it was red?

Once the idea of conditional probability is clear, the Bertrand box paradox can be described. Three boxes are given, each with two drawers. There is a coin, gold (G) or silver (S), in each drawer. One box has a gold coin in each drawer (GG), another box has a silver coin in each drawer (SS), and the third box has a gold coin in one drawer and a silver coin in the other drawer (GS). A box is chosen at random, a random drawer in the box is opened, and a gold coin is found inside it. What is the probability of the coin in the other drawer being gold?

The following, wrong, answer is commonly given. Initially, each of the three boxes had the a probability of 1/3 to be chosen, $P(GG) = P(SS) = P(GS) = 1/3$. The box actually chosen had gold, so could not be box SS; it must have been GG or GS. These two boxes are now equally likely to be chosen. The only box where the other coin is also gold is GG, which is why the desired probability is 1/2.

This result is wrong for the (subtle) reason that the phrase "two boxes are now equally likely to be chosen" is wrong. Initially, the two cases GG and GS were equally likely, with probability 1/3 each. Now, however, we know that SS was chosen, so this becomes a case of conditional probability. If box GG will be chosen, the probability of finding a gold coin in its other drawer is 1, while if box GS will be chosen, the probability of finding a gold coin in its other drawer is 0.

To find the correct probability we apply Bayes' rule, which in our case becomes

$$\begin{aligned} &P(GG|\text{found gold}) \\ &= \frac{P(\text{found gold}|GG) \times \frac{1}{3}}{P(\text{found gold}|GG) \times \frac{1}{3} + P(\text{found gold}|SS) \times \frac{1}{3} + P(\text{found gold}|GS) \times \frac{1}{3}} \\ &= \frac{1}{1+0+1/2} \\ &= \frac{2}{3}. \end{aligned}$$

◇ **Exercise 6.10:** A woman is walking with a boy. She drops her bag and we pick it up and return it, which starts a conversation. After thanking us, she tells us that she has two children and that the boy is her son. What is the probability that her other child is also a boy? (See reference [Bar-Hillel 82] for this and other examples of conditional probability.)

• The Monty Hall problem. (This problem or paradox is listed here even though it is described in a vast number of sources. I decided to discuss it because (1) It is probabilistic and it illustrates how probability problems can be subtle and exceedingly unintuitive. (2) It is often explained in complex, and opaque statistical terms, but the explanation is simple and easy to understand and accept. (3) It gave me a chance to include several quotations that I like.) I chose [WikiMonty 23] as one reference out of the many that are available.



The Monty Hall problem is based on the American television game show *Let's Make a Deal* and is named after its original host, Monty Hall. It was originally published in the scientific journal *American Statistician* in 1975, and in 1990 became famous when it appeared in *Parade magazine* in the column "Ask Marilyn." This column was written by Marilyn vos Savant, who has also solved it, to much discussion, opposition, and war of words. (Marilyn vos Savant is an American magazine columnist who has the highest recorded intelligence quotient (IQ) in the *Guinness Book of Records*.) The problem is easy to explain and to grasp.

> Hi Marilyn, wherever you are.
>
> —Arthur C. Clarke about his ex wife Marilyn Mayfield, in *The View from Serendip*, (1977), New York: Random House.

The problem is easy to state and understand. Participating in a game show, you are given the choice of three doors: Behind one door is a car; behind the others there are goats (Figure 6.10). You pick a door, say #1, and the host, who knows what is behind the doors, opens another door, say #3, which has a goat. He then says to you, "Do you want to change your mind and pick door #2 instead of #1?" Is it to your advantage to switch your choice?

Most people who considered this problem, among them logicians, probabilists, and statisticians, argued that once you have eliminated door #3, the problem reduces to just two unknown doors, which now have equal probabilities. It was the genius of Marilyn vos Savant to notice the subtle point that changes the situation, namely that the contestant does not know what is behind the doors, but the host knows! It turns out that this knowledge changes the probabilities from 0.5 and 0.5 to 1/3 and 2/3 in favor of door #2, i.e., the contestant should change his or her mind. Here is the explanation that I believe is the simplest.

Instead of three doors, imagine five doors. Initially, each door has a 0.2 probability of having the prize. After the contestant chooses, say, door #1, the remaining four doors have a collective probability of 0.8. The host, knowing where the prize is, opens, say, doors 2, 3, and 4, revealing goats. At this point, the 0.8 probability has completely



Figure 6.10: Image by MS Copilot, Aug. 2025.

transferred to door #5, making the final choice obvious.

> Obvious is the most dangerous word in mathematics.
> —Eric Temple Bell.

This explanation becomes even more convincing if we consider a million doors instead of just three or five. Each door now has a probability of 0.000001, so the initial choice of a door is very likely wrong. The host opens 999,998 doors, revealing goats, which changes the probability of the last remaining closed door to 0.999999, extremely large.

Table 6.11 illustrates how the two choices made by the game show player leads to either a win or a loss and the probabilities of these events. The bottom part of the table is a summary of the nine possible results in the case of three doors, and it is obvious that switching the doors doubles the chance of winning.

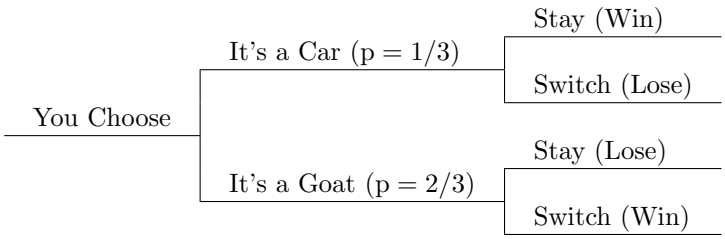
Enough with the goats, they smell like a barn and are constantly chewing something.

- Berry's Paradox. Consider the following sentence "The least integer not describable in fewer than twenty syllables." This sentence consists of nineteen syllables, which implies that the—least integer not describable in fewer than twenty syllables—is describable in fewer than twenty syllables; a paradox. See also Exercise 2.9.

The simple concept of a successor offers a possible resolution. Any natural number (nonnegative integer) can be described as the successor of the one preceding it, which makes it possible to write a sequence such as:

0 is a number.
1 is the successor of the number defined in the last line.
2 is the successor of the number defined in the last line.
3 is the successor of the number defined in the last line.

Assuming that the digits count as one syllable, each of the lines above consists of 16 syllables, which shows that there is no smallest number that can be described in no fewer than twenty syllables. Once we agree with this, Berry's Paradox disappears, but there are reasons to object to this explanation. (1) The digits must be spelled, leading to lines with more and more syllables, and (2) the description of the number n must



Car located	You choose	Monty opens	Stay	Switch
Door1	Door1	Door2 or Door 3	Win	Lose
Door1	Door2	Door3	Lose	Win
Door1	Door3	Door2	Lose	Win
Door2	Door1	Door3	Lose	Win
Door2	Door2	Door1 or Door 3	Win	Lose
Door2	Door3	Door1	Lose	Win
Door3	Door1	Door2	Lose	Win
Door3	Door2	Door1	Lose	Win
Door3	Door3	Door1 or Door 2	Win	Lose

Table 6.11: Summary for Three Doors.

include the n lines that precede its line. Well, perhaps this paradox will stay unresolved for now (sigh).

- You are now invited to use your gray cells, à la Hercule Poirot, to try to grasp the sophisticated Grelling-Nelson paradox. This 1908 antinomy is a semantic self-referential paradox, due to Kurt Grelling and Leonard Nelson, which asks whether the word heterological applies to itself. First, some terms. Autological (also called homological) is a word that describes itself. The word "pronounceable" is pronounceable, "polysyllabic" consists of several syllables and is therefore polysyllabic. The word "noun" is itself a noun; it is not a verb, so it is autological. In contrast, the term heterological refers to a word that does not describe itself. Examples are "green" (it is not green), "circle" (it is not circular), and "hyphenated" (which I hope is not hyphenated).

  The paradox is the simple question: is the word "heterological" itself heterological?
- An answer of no implies that "heterological" describes itself and is therefore autological. Since it is autological, it must describe itself and is therefore heterological. A contradiction!
- A yes answer implies that the word "heterological" is itself heterological and therefore should not describe itself. But it does describe itself, because it says, in effect "I am heterological." Therefore it is not heterological. An impasse!

Enough is enough. Such long words and convoluted logics can give even a whale a substantial headache. (But see elsewhere in this book, especially the Ross-Littlewood paradox on Page 249.)

However, lovers of paradoxes should consult the long list found at [WikiParadoxes 24].

6. Paradoxes

How wonderful that we have met with a paradox.
Now we have some hope of making progress.

—Niels Bohr.



7
Probabilities: the Rule of Chance

One of the many branches of mathematics is probability, the study and measure of the likelihood of events and the truth of propositions. We learn intuitively the meaning of probability from our life experiences, but in order to turn this vague term into a practical and important concept, it has to be quantified, which allows for probability calculations and estimations.

Here is a sample that illustrates the significance of probability and luck. There is a raffle in which only one ticket will be selected as the winner. One million tickets are sold rapidly and you purchase the ticket with the number 32,768. You reasonably assume you won't win, as there is nothing distinctive about your number. If you think you won't succeed, you will feel that way no matter what your specific chances are. However, you are aware that one number is certain to emerge victorious. Therefore, your belief and your knowledge do not align, creating a paradox. Introducing the concept of probability solves this straightforward paradox. There is no certainty that you will lose. You are aware that your chances of winning are minimal, but this does not go against your understanding that someone will emerge as the winner.

7.1 Basic Concepts

The first step in quantifying probability is to locate (or to generate artificially) a set of independent, equiprobable events, to choose an outcome A, and to count how many events end with A. (The important term "independent" is discussed a few paragraphs down.) The probability of outcome A is then denoted by $P(A)$ and is defined as the ratio

$$P(A) = \frac{\text{number of cases with outcome } A}{\text{total number of cases}}. \tag{7.1}$$

Equation (7.1) implies the following basic features of a probability:

- A probability is always non-negative.
- Of outcome A never occurs, then $P(A) = 0$, and if it always happens then $P(A) = 1$.
- The probability of a result A not occurring is the difference $\overline{A} = 1 - A$.
- The special case of outcomes A and B never occurring in a set of events is given by $P(A \text{ or } B) = P(A) + P(B)$, but this is a special case. The general case is given by Equation (7.2).

Thus, probabilities are always numbers in the interval $[0, 1]$, where the numeric values in this interval lead from impossibility to certainty. However, the discussion of probability in this chapter and the many examples and surprises shown here serve to emphasize the fact that probability is a subtle concept, full of surprises, unintuitive results, and conceptual pitfalls.

> "Great." He holds my gaze for a beat. "Bella's going flip when I tell her I ran into you. What are the odds?"
> "In a city of nine million? Less than zero."
>
> —Rebecca Serle, *In Five Years*, 2020.

One such pitfall has to do with the two extreme values, 0 and 1, of a probability. In everyday speak we assume that 0 probability indicates impossibility of an event, while 1 indicates certainty, but consider an event with an infinite number of possible outcomes, such as randomly turning a wheel of fortune. The result is the angle of turn, a real number in the interval $[0, 360]$, that we divide by 360 to obtain a real number in the range $[0, 1]$. The probability of obtaining any given real number, such as 0.8604263, is zero. In practice we may not even be able to measure such a result accurately on a wheel or in a computer, but most of the time we are satisfied with a close approximation, such as 0.86. Thus, when a result such as 0.00003 is obtained, we may replace it with 0 and realize that 0 means a result very close to 0. The event in question may not happen in our lifetime, but it may happen in principle.

A similar surprise is the percentage of transcendental numbers among the real numbers. This is well described by the statement, found on Page 129 "If you select a real number at random among all the real numbers, it will be transcendental with probability 1!"

There are areas of science, dealing with very large systems, in which probability methods and techniques must be used instead of exact treatment. Perhaps the simplest example is gas, which involves dealing with vast numbers of molecules. The atomic weight of oxygen is 16, which is why 16 grams of oxygen contain $\approx 6.02\times10^{23}$ (Avogadro's number) atoms of oxygen. (Free oxygen exists in the form of molecules, which contain two atoms each, so the number of oxygen molecules is half that of atoms.) A precise, deterministic treatment of such vast numbers is out of the reach of existing technology, which is why probabilistic (or rather statistical) description of the properties of real gases is feasible. A similar example is weather forecasting, an important science which deals with large numbers of variables that change constantly.

Another common mistake we tend to make when computing probabilities is handling partitions. This is easy to understand when we consider dice throwing. A die has six

sides, which is why throwing a single die results in one of six outcomes. Throwing two dice produces one of $6^2 = 36$ results, and in general, there are 6^n possible outcomes when throwing n dice. The case of $n = 3$ is simple enough to consider and analyze in detail. There are $6^3 = 216$ possible outcomes, but they are not equiprobable. The smallest result, 3, is obtained with $(1, 1, 1)$, and the largest outcome is a throw of $(6, 6, 6)$, but an outcome of 9 is obtained in the six cases $(6, 2, 1)$, $(4, 3, 2)$, $(3, 3, 3)$, $(2, 2, 5)$, $(1, 3, 5)$, and $(1, 4, 4)$, which raises the question of how many times to count each of these partitions in the calculation of the probability $P(9)$.

The correct answer, already understood by Galileo, is not immediately obvious and depends on the number of different components in a partition. The simple rule is as follows:

- A partition of the form (x, x, x) should be counted once as an outcome for the purpose of computing a probability.
- A partition of the form (x, y, x) should be counted three times, as (x, y, x), (y, x, x), and (x, x, y). This is because its two equal components cannot be distinguished.
- Similarly, A partition of the form (x, y, z) should be counted $3! = 6$ times because its components can be permuted this many times.

In the case of outcome 9, for example, partition $(3, 3, 3)$ should be included once, partitions $(2, 2, 5)$ and $(1, 4, 4)$ should each be included three times, and partitions $(6, 2, 1)$, $(4, 3, 2)$, and $(1, 3, 5)$ should each be included six times. The probability of obtaining 9 in the throw of three dice is therefore

$$P(9) = \frac{1 + 2 \times 3 + 3 \times 6}{216} = \frac{25}{216} \approx 0.1157,$$

or approximately 11.6%.

Probability may also be of interest to gamblers, philosophers, poets, medical workers, and others, but mathematicians like to treat this concept rigorously, which is why an entire branch of mathematics, probability theory, has been developed and is continually studied and expanded, in order to place the nebulous, intuitive concept of probability on a firm mathematical foundation. Such a process starts with a set of axioms (postulates or assumptions that are self evident) on which are built new and useful terms, theorems, and concepts.

The chief terms used in probability theory are an experiment, event, probability measure, and sample space. An experiment is a test that involves real, physical objects or quantities. It can be real or a thought experiment. Examples of simple experiments that are common in the study of probability are the throw of a die or a coin, choosing a card from a deck of playing cards, and turning a wheel (of fortune or of a roulette). An event is the result (outcome) of an experiment. In the case of a coin, the two possible events are head and tail. The sample space of an experiment is the set of all possible events of the experiment. A probability measure is the probability of an event.

> ...the view that the universe is essentially probabilistic in character, or, to put it in more colloquial language, that the world is full of random happenings.
>
> —Patrick Suppes, *Probabilistic Metaphysics*, Oxford: Blackwell, 1984.

Probability calculations. Another unusual aspect of mathematics is found in probability calculations, which may be very subtle. First, a short discussion of the difference between probability theory and the science of statistics. These two disciplines are in some sense opposites or inverses of each other. In probability, we start with a process driven by randomness or uncertainty, and we try to figure out the results of running it. It is like predicting the future. In statistics, we start with known past events and then try to guess the process that generated them or that can explain them.

The events dealt with in probability calculations may be dependent or independent. Snowy days are a good example of dependent events. The event "it is snowing today" affects the probability that it will snow tomorrow. On the other hand, a house fire that happens today in our neighborhood does not affect (we hope) the probability that another house fire will happen in a different neighborhood tomorrow. These events are independent.

The probability that two independent events will occur is simply the product of the individual probabilities. Thus, $P(A \text{ and } B) = P(A) \times P(B)$. Perhaps the simplest example is the probability of getting heads in two consecutive throws of a coin. The probability of getting a head in each throw is $1/2$, so the combined probabilities is the product $1/2 \times 1/2 = 1/4$. It is smaller than each of the individual probabilities, which brings us to the important fact that a probability must always be a number between zero and 1. Computations with probabilities must therefore be done with care, to keep any results within this interval. It is therefore wrong to simply add probabilities without subtracting anything, because this may result in a sum greater than 1. In general, probability computations involve operations such as $A \times B$ and $1 - A = \overline{A}$.

Example: Dependent events. It is easy to see why the following two events are dependent. Simply multiplying probabilities of those events produces absurd results.

Event A: Only 1% of the population of Shambhala are taller than 190 cm (six feet three inches). If you pick a random Shambhala population of 10,000, you can expect about 100 to be that tall.

Event B: Only 500 people of Shambhala's population of 500 million played professional basketball in year 2000. The percentage in this case is 10^{-6}.

What is the probability of a Shambhala resident to be both tall and a professional basketball player? Simply multiplying $10^{-6} \cdot 10^{-1}$ produces 10^{-7}. This low value implies that we can expect only $5 \cdot 10^6 \times 10^{-7} = 0.5$ people to be in this category. However, a short visit to a Shambhala training camp for basketball players verifies that perhaps 80–90% of them are tall.

Assuming that we know how to compute the probability of A and B (the product of two events), how do we compute that of (A or B), the so-called sum of two independent events. The trick is to start with the inverse probabilities $\overline{A} = 1 - A$ and $\overline{B} = 1 - B$, and examine a new case C where both inverse events occur, i.e., A does not happen *and* B does not happen. The probability of C is therefore the product $P(\overline{A}) \times P(\overline{B})$. Now consider the inverse of C. This is the case where both inverse events do not occur, i.e., either one *or* the other occurs. Thus, the sum (A or B) of two events is the inverse of C, implying that its probability is

$$P(A \text{ or } B) = 1 - P(C) = 1 - \big(1 - P(A)\big)\big(1 - P(B)\big). \tag{7.2}$$

It's easy to see why Equation (7.2) makes sense. Given two events A and B with high probabilities of occurrence, it is clear that $P(A \text{ or } B)$ must be high, and this is also implied by the equation. If $P(A)$ is large, then $(1 - P(A))$ is small, and similarly for $P(B)$, which produces a large value for $P(A \text{ or } B)$.

If the probabilities of A and B are 0.65 and 0.95, respectively, then the probability that either A or B will happen is $1-(1-0.65)(1-0.95) = 1-0.35\times 0.05 = 0.9825$, much higher than either of the individual probabilities. This simple example of determining the probability of the sum of two events illustrates how computing the probabilities of even simple events may be subtle and far from trivial.

> It is always probable that something improbable will happen.
> —Logan E. Bleckley.

7.2 More Probability Concepts

A Bernoulli trial is a random experiment or event with two possible outcomes that are normally termed success and failure (or yes and no, 1 and 0, or anything else) and where the probability of success is constant (implying that the probability of failure is also constant).

Expected value. Imagine a random experiment that is run (or an event that occurs) many times, resulting in different outcomes. In the simplest case, where the outcomes occur with the same probability, the expected value of the experiment is the arithmetic average of the outcomes. In the general case, where an outcome x_i occurs with probability p_i, the expected value of the experiment is the weighted average of the outcomes, $\sum_i x_i p_i$. In practice, we count the number of times w_i each outcome x_i occurs and we can either compute the individual probabilities $p_i = w_i / \sum_k w_k$ or compute the expected value of the experiment as the weighted sum

$$\frac{\sum_i x_i w_i}{\sum_i w_i}.$$

Notice that the value of an arithmetic average or a weighted sum may be different from any of the individual outcomes x_i. Thus, the expected value of an experiment may not be any of its actual outcomes. Perhaps the simplest example is a dice throw. Throwing a die results in a value between 1 and 6, but the expected value of such an experiment, the average, is 3.5.

Probability and likelihood. Experimental scientists, such as physicists and biologists, perform experiments. They observe and record phenomena and draw conclusions about unknown properties of the world around us. We can say that those people deal with likelihoods. They can only say that a certain outcome of an experiment is, say, 80% likely. Theoretical scientists, on the other hand, propose various properties and calculate the probabilities of what we can expect to observe if those properties were real.

In general, if the outcomes of an experiments are known in advance, we refer to them as probabilities. In order fo them to make sense they must add up to 1, indicating that

some outcome must occur. If the outcomes are unknown, we can perform experiments and assign likelihoods, not probabilities, to the observed outcomes.

A simple practical example is a coin throw. If the coin is fair, we know the probabilities of the outcomes in advance, and we also know that one of the two outcomes must happen. In contrast, if the coin is unfair, the only conclusion we can draw from an experiment (the throw of such a die) is a likelihood. We don't know how much the coin is unfair, so we can only observe outcomes and assign likelihoods, because we don't fully understand the factors that determine the results of the experiment. Given enough time and experiments, those likelihoods may converge to probabilities.

In other words, if we throw a coin and observe an outmode of heads, we don't know whether the coin is fair or unfair, so we can only make an observation and assign a likelihood to this outcome.

Law of large numbers (LLN). We know from experience that any individual outcome of a random experiment may be far from the expected value of the experiment, but we also know, or trust, that repeating the experiment many times will result in outcomes whose average (arithmetic or weighted) will get closer to the expected value. The law of large numbers guarantees this behavior.

A practical example is a winning streak. It sometimes happens that a gambler in a casino experiences a series of consecutive (or almost consecutive) wins, which encourages him to continue to play, hoping for his luck to hold. Such behavior is known as a gambler's fallacy. The LLN, however, dashes this hope and guarantees that in the long run, the gambler would lose and the casino would win, according with the expected value of the particular game being played.



Figure 7.1: Image by MS Copilot, Aug. 2025.

The LLN is named after Jacob Bernoulli and the phrase "large numbers" refers to the large number of trials needed for the result to approach the expected value. There are random experiments whose expected values are infinite or do not exist, and such cases do not obey the LLN.

Law of truly large numbers. This is more a proverb than a law. In plain words it states that given a very low-probability (an extremely unlikely) event x, the probability

of it occurring in n trials can become arbitrarily close to 1, if n is truly large. This is true regardless of how small the probability $p(x)$ is. An important consequence of this law is that even very unlikely events may affect the overall behavior of a random experiment, and should therefore not be ignored without further consideration. Here are two examples that the law may explain.

- Psychics. History has many examples of people claiming to predict the future. Such a person makes predictions of future events and publicizes them. The law proposes that the more predictions are being made, the more would prove correct, and because of the gullibility of people, the few correct predictions would tend to erase the memory of the many wrong ones.

- Similarly, when a gambler wins, this low-probability event helps him forget the many high-probability losses he had suffered. This is termed selective memory bias.

Elementary probability calculations show why we should accept this law. Given an event with probability p, the probability that it doesn't happen is $1-p$, and the probability that it never happens in n trials is $(1-p)^n$. Thus, the opposite of this event, i.e., the probability that is does happen at least once, is $1-(1-p)^n$. It is easy to see that if p is small, then $1-p$ is close to 1, causing $(1-p)^n$ to be smaller and smaller as n gets bigger. This, in turn, causes $1-(1-p)^n$ to get bigger and approach 1 for very large values of n.

For $p = 0.1$, 1000 independent trials produce $(1-p)^{1000} \approx 0.3677$, which implies that the probability of such a low-probability event to happen during 1000 trials is $1 - 0.3677 = 0.6323$, quite high.

For the very low value of $p = 10^{-7}$, similar calculations produce

$$1 - 0.999999999^{1000000000} \approx 0.63,$$

A surprisingly large value.

A close relative of the law of truly large numbers is the black swan theory, reference [BlackSwan 24]. It deals with real, historical events that were extreme outliers; rare, unexpected, and important. The theory argues that such events, even though rare, and even though rationalized by hindsight, played important roles in history and affected its progress more than the many familiar, common historical milestones.

Another cousin of the law of truly large numbers is Littlewood's law which states that a person can expect to experience events with odds of one in a million at the rate of about one per month. This intriguing observation is due to J. E. Littlewood who developed it in an attempt to curb and debunk claims of supernatural phenomenology.

The quantitative justification of Littlewood's law assumes that an average person experiences an event (either trivial or significant) every second and does that for eight hours a day. This rate translates to $60 \times 60 \times 8 \times 35 = 1008000$ events every 35 days, implying that it is reasonable to expect one rare event, with probability 10^{-6}, about once a month. Thus, rare events should be common and should happen more often than we intuitively expect.

◇ **Exercise 7.1:** (A joke). What is the law of small numbers?

> In fact, my mother was more mathematically inclined than she realized. She used probability every day of her life and called it intuition.
>
> —Liane Moriarty, *Here One Moment*, Sept. 2024.

7.3 Randomness

Determinism is the philosophical view that events are completely determined by previously existing causes. Causal determinism is the idea that every event—even our deliberations, choices, and actions—is necessitated by past events and conditions, working together with the laws of nature. However, the occurrence or existence of yet other things depends upon our deliberating, choosing, and acting in a certain way.

Our view of randomness has varied considerably over time, swinging from determinism to randomness and chaos. In 1814, Pierre-Simon de Laplace has published his belief in causal determinism. Essentially, he stated that if an intellect exists which, at a certain moment knows all forces that set nature in motion, and all positions of all items of which nature is composed, he (or it) would embrace in a single formula the movements of the objects, from astronomical bodies to the tiniest atoms. For such an intellect, randomness would not exist. Nothing would be uncertain and the future would be completely predictable and pre-determined.

Randomness

Chaos weaves its tapestry, a dance untamed,
Random whispers in a wild, wild game.
No pattern guides the cosmic dice,
Chance reigns supreme, a gamble precise.

In quantum realms, uncertainty's delight,
Particles waltz in unpredictable flight.
A roll of dice, a shuffled deck,
Randomness unveils the unforeseen trek.

Nature's brush strokes, scattered and free,
A cosmic canvas in random decree.
Uncharted paths, where chaos gleams,
In randomness, life births its dreams.

Composed by ChatGPT Jan 2024.

In 1867, James Clerk Maxwell had a similar idea which today is known as Maxwell's demon. The nature of heat as movements of atoms and molecules became clear in his time, so Maxwell imagined an intelligent being that can see molecules. By allowing only fast molecules into one side of a container and only slow molecules into the other side, such a being could "magically" create hot and cold parts in the same container, an effect that contradicts everything we see in nature. This was an attempt to convert a random system into a well-organized one without the use of external energy.

Once radioactivity was discovered, in 1896, early researchers in this field quickly realized that radioactive decay involves randomness. It is impossible to predict when a

given atom would decay. Only the probability of its decaying in the next time unit is a meaningful, physical quantity that can be measured and used in experiments. This way of thinking has reintroduced randomness into science, a process that continued in the early 1920's by the creators of quantum theory.

Later in the 20th century, mathematicians started gaining deeper understanding of randomness and its relation to determinism, and today we feel that the behavior of quantum particles is random, but aggregates of the same particles exhibit behaviors that look more deterministic as more particles are involved.

Faced with something that seems to have no patterns and looks unpredictable, we tend to assign it the label "random," but this term is only an intuitive, imprecise name for something that is very common, that seems to pervade the world around us, and that should be studied further, understood better, and be defined more precisely.

The term "random" is commonly but wrongly used (misused) in many contexts that involve rare events. Phrases such as "I was randomly walking when..." and "She randomly ran into her ex at the gym," are common. Such events may be rare and unrelated, but are not the same as a patternless process.

Randomness is an important concept that is closely related to and much used in probability studies and work. A simple, working definition of randomness is the absence of a pattern, but randomness is more than that. Here are some thoughts to help explain and shed more light into this important concept:

> Life is a school of probability.
> —Walter Bagehot.

- Many processes that seem random to us are actually deterministic, but there are processes that are (or at least we currently believe that they are) truly random. The familiar trials of coin flips (tossing a coin or flipping it on its edge on a surface) and dice rolls seem random, but are in fact deterministic [Keller 86]. An entity who knows the initial conditions and the forces and velocities that affect a toss or a roll could in principle compute the final outcome in advance, and be correct every time! See also Section 7.5.

> Rival captains James Rodriguez and Justo Villar were left scratching their heads over a bizarre coin toss before Colombia's Copa America clash with Paraguay.
> Rodriguez made the call as the referee flipped the coin only for the U.S. nickel to land perfectly on its edge, wedging in the grass on its side.
> The referee threw up his hands in amazement, while Rodriguez and Villar both laughed out loud.
> The odds of a coin, specifically a U.S. nickel, landing on its edge are 6000:1.
> —Yahoo Sport Australia, 8 June 2016.

It is possible, even practical, to construct a coin flipping machine that consistently applies the same force to a coin and throws it in the same direction and at the same orientation, such that it always lands heads. Therefore, coin tosses, which are random in practice, are not random in principle. They are completely determined by physics causality, and their apparent randomness is the result of the input variables that often vary widely. We use coin tossing often to demonstrate concepts of probability and randomness, and yet this simple and well-known process is physics, not random. In

The Coin That Shouldn't Stand

Javier and Mateo grew up in neighboring towns in the Andes, both obsessed with football and probability. Javier, a goalkeeper with lightning reflexes, studied physics in his spare time, fascinated by chaos theory and the geometry of flight. Mateo, a midfielder known for his precision passes, had a secret love for street magic and sleight of hand.

They met at a regional tournament in Quito, where their teams clashed in a muddy semifinal. After the match, they bonded over a shared superstition: the belief that improbable events, like a coin landing on its edge, were signs from the universe. They called them "moments of the hinge," when reality briefly opens to something deeper.

Years later, now semi-retired and coaching youth teams, they traveled together to a sports psychology conference in California. During a break, they wandered outside, still in their team jackets, and began debating whether luck could be trained. Javier pulled a nickel from his pocket and flipped it, joking, "If it lands on its edge, we'll know."

It did.

They froze. The coin stood upright on the rough ground, defying gravity and probability. For a moment, neither spoke. Then Mateo whispered, "This is one of those hinges."



It's ready now, Javier and Mateo are no longer just two men pondering a coin; they are standing on the edge of probability in a packed stadium, where every seat and shadow adds weight to the moment. The crowd, the architecture, the tension in the air, it all amplifies the surreal hinge of that nickel standing upright.

—Image and story by MS Copilot, Sept. 2025.

practice, we make it seem random by tossing it with slightly different forces and directions. Since its movement in the air is described by nonlinear expressions, any small change in its initial conditions is greatly amplified during its short trip, which causes a large uncertainty in the result (see Page 356).

(The following bias is known from experience. A coin tossed and caught by hand is

slightly fairer than a coin tossed and allowed to end up on the floor. This is because a coin hitting the floor tends to spin before it comes to rest.)

On the other hand, some processes are currently believed to be truly random, and here are three examples:

1. Quantum mechanics claims that many processes on the atomic level are random and all we can ever hope to achieve is to discover or compute their probabilities. This is because of the nature of our universe and not because of our limited knowledge or lack of computing power. One such process is radioactive decay, and reference [hotbits 24] explains how to use it to generate random bits.

2. The movements of gas molecules are believed to be truly random, which is why URL `random.org` uses atmospheric noise to generate and freely provide strings of random bits.

3. Similarly, the "snow" that used to be so common in old television sets was believed to be random, but I don't know of any attempts to use it to generate random strings.

4. Reference [Kac 84] describes an elegant experiment where a small mirror suspended from a quartz fiber in a tank full of stable air (no air currents) was discovered to rotate randomly because of collisions with air molecules. The number of collisions per second is vast, so we assume that at any time, the mirror is bombarded equally from all sides, but the variance in the speed of the individual molecules is great, which occasionally results in a pressure difference that pushes the mirror slightly, causes it to randomly rotate on the thread. This is similar to the effect of Brownian motion described on Page 358.

Albert Einstein is known to have claimed that God does not play dice with the world. Today we may say that if he was wrong, if God does play dice, then His are truly random, perfect dice that do not depend on our physics.

- Irrational numbers have been discussed in Section 2.4. Such a number has an infinite number of digits to the right of the decimal point, and this sequence of digits is random. It has no pattern and it is impossible to guess the next digit in this sequence even after examining all its predecessors. However, such a sequence is deterministic and it is possible to compute any number of its digits, given enough time and computer storage.

> Is randomness random?
> —An unanswered question.

- The outcomes of an experiment may be random, but may be subject to a probability distribution. A throw of a pair of dice results in an integer between 2 and 12. The next outcome of such a throw is unpredictable and therefore random, but repeating this experiment many times shows that some integers in this interval will be more common than others. The sequence of dice throws is random, but the distribution of its results is not.

- Random numbers are used in probability and other computations, which is why it is important to be able to generate long random sequences of numbers by software. However, software is deterministic. There is nothing random about the results generated by a computer program. Therefore, the random sequences of numbers generated by

software are, in principle, not random. They appear random, they satisfy the statistical tests for randomness, which is why they are termed pseudo-random.

- Computer users may sometimes try to generate a random sequence of characters by randomly hitting keys on a computer keyboard, but the results of such a process are far from random. One reason is that as we quickly touch keys, we tend to use our hands alternately. The result is a sequence where consecutive symbols tend to come from different sides of the keyboard. Another reason for the non-randomness of such a sequence is the frequency of occurrence of double and triple symbols. We intuitively feel that two consecutive occurrences of the same symbol would be rare in a random sequence, but it is easy to show that in a sequence of decimal digits, where there are only 10 different symbols, the probability of a double is $0.1 \times 0.1 = 0.01$ or 1%, not very rare.

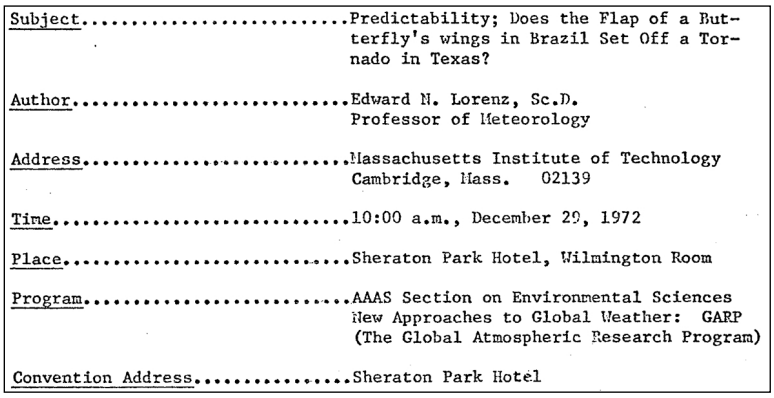
- Humans in general, not just computer users, are a poor source of random numbers. Reference [37everywhere 24] shows that when we are asked to produce a random integer between 1 and 100, the most common choices are 3, 37, and 73. See Section 7.10 for an unexpected appearance of 37 in real life.

- Linear computational processes are deterministic and generate results that are reproducible. When we compute consecutive elements of the Fibonacci sequence, which is defined by $x_{n+2} = x_{n+1} + x_n$, we always get the same sequence. The sequence, however, depends on the initial conditions, which in the Fibonacci sequence are the values of x_1 and x_2. Subsections 5.6 through 5.6.2 show how changing those conditions produces different sequences, but the point is that varying the initial conditions of a linear process just slightly, varies the final results only a little. This property of linear processes makes them useful in many areas of science and mathematics, because a user can perform a series of experiments while varying the initial values in small steps, and the results would also vary in small steps and in a predictable way.

  Nonlinear computational processes behave in a completely different way because they tend to be sensitive to initial conditions. Such a process can be deterministic and can generate results that are reproducible and do not depend on chance in any way. However, such computations tend to generate results that are very sensitive to the initial conditions. Running such a process many times, while changing the initial conditions only slightly, may produce results that look and behave completely differently, and may thereby create the illusion that the process includes an element of randomness.

  A common example of a nonlinear process is the simple sequence produced by the rule $x_{n+1} = r \cdot x_n(1 - x_n)$. This is a nonlinear rule of computation because of the term $-x_n^2$. Its only initial value is x_1, and it is easy to run for many values of x_1 and examine the behavior of the resulting sequence. Of the many references that describe and illustrate the unexpected and unusual behavior of this sequence, only [WikiBifurcat 23] and [Bifurcation 20] are mentioned here.

  This surprising, and perhaps even disturbing, behavior of nonlinear processes was discovered in 1961 by the meteorologist Edward Lorenz while working with early computers, trying to forecast weather based on a complex mathematical model. Later workers in the field of mathematical chaos have extended his work, understood nonlinear processes better, and created the well-known adage (Figure 7.2) "Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" See also Page 417.



Subject	Predictability; Does the Flap of a Butterfly's wings in Brazil Set Off a Tornado in Texas?
Author	Edward N. Lorenz, Sc.D. Professor of Meteorology
Address	Massachusetts Institute of Technology Cambridge, Mass. 02139
Time	10:00 a.m., December 29, 1972
Place	Sheraton Park Hotel, Wilmington Room
Program	AAAS Section on Environmental Sciences New Approaches to Global Weather: GARP (The Global Atmospheric Research Program)
Convention Address	Sheraton Park Hotel

Figure 7.2: The 1972 Announcement of Chaos.

◇ **Exercise 7.2:** Chaos theory is now well established and there are many sources and examples of nonlinear processes whose unpredictable outcome may seriously affect our future. Find such an example.

■ Another experiment that further blurs the boundary between randomness and determinism is described by Mark Kac in Chapter 3 of reference [Kac 85] as well as in [Kac 84]. There he shows that given any truly random sequence R, it is possible to deterministically compute another sequence D that would not be identical to R, but would be similar enough to make it impossible to distinguish it from R by means of any statistical test for randomness. It would then be impossible to tell which sequence is the "true" random one and which is the artificial, deterministic one. Kac then finishes by asking "So, what is chance?"

> In English, the phrases "a fat chance" and "slim chance" have the same meaning.
> —Anonymous.

We generally agree that the best way to generate a truly random sequence is by sampling certain natural phenomena such as the times between consecutive radioactive decays, Brownian motion of a small particle, the "snow" sometimes seen on television screen, and the fluctuations of the stock market index. However, Kac's claim is that no matter how "natural" the source of a random sequence is, it can always be mimicked by a deterministically-generated sequence that would not be identical to, but would be indistinguishable from, the original sequence. This challenges our intuitive idea that there is absolute randomness and that randomness and determinism are opposites.

■ Considering the points above, perhaps we can claim that chance is a fractal. Fractal, from Latin fractus (broken), is a word coined and popularized by Benoît Mandelbrot. It is a geometric shape containing detailed structure at arbitrarily small scales. Many fractals appear similar to themselves at smaller and smaller scales, and are accordingly referred to as self-similar. A curve is normally thought of as being one-dimensional (this is its topological dimension) and a surface, similarly is two-dimensional. But fractal curves can be assigned fractal dimensions in the interval $[1, 2)$ and fractal surfaces can

similarly have dimensions greater than 2. An example is the random path taken by a small particle which is subject to Brownian motion. It has been conjectured, and then proved, that the fractal dimension of such a curve is 4/3, thereby relating randomness and chance, which are probabilistic concepts, to fractals, which are geometric figures.

■ Brownian motion is caused by the fast movements and many collisions of gas or liquid molecules. When such molecules hit a small object such as a pollen particle, they cause it to move erratically in a path that looks random.

The following numbers serve to put the scale of those processes into focus. Unlike collisions between macroscopic objects, collisions between microscopic particles are elastic; there is no loss of kinetic energy. A molecule in gas at room temperature can suffer on the order of 10^{14} collisions per second from other molecules. Gas molecules at room temperature are moving at up to 300–400 meters per second, but there is a huge variation in speed. While some molecules move fast, other may move very slow, and it is this wide speed variation that creates the Brownian effect. It often happens that more fast molecules hit a pollen particle from one side and slow molecules hit it from the opposite side.

However, some may claim that the randomness of Brownian motion is not "true," because even though molecules are microscopic particles, their movements are governed by the deterministic rules of classical physics. The randomness that we perceive is the result of the vast numbers of molecules and the large variance in their speeds. If we accept this argument, then the phrase "a pollen particle is hit by gas molecules at random," simply hides the fact that the randomness is illusory and is the result of the large numbers involved.

■ Randomness and information. We are familiar with the term "information." We hear all the time that information is power, that we live in the Information Age, and that more and more people are information workers, information technicians, or information manipulators. However, like so many other basic concepts and terms, information is difficult to define rigorously. We intuitively feel that information is about order. A long string of random text conveys no information, whereas a short string of familiar words may have important meaning if the words are ordered in a certain way.

On the other hand, it has long ago been observed that the distribution of letters in typical text is far from uniform. Today it is easy to scan, say, the complete works of Shakespeare and discover that the letters `E`, `T`, and `A` are the most common in this collection of fine English, while `x`, `j`, `q`, and `z` are the rarest. The pseudo words `ETAOIN SHRDLU` have come to denote this property of natural languages.

A curiosity. In English, as in various other languages, words may contain silent letters, letters that are normally associated with sounds, but are silent in special cases. The following list, due to YouTube channel `Vsauce`, is interesting because it lacks the letter `V`. This is the only letter which is never silent in English pronunciation.

BREAD, CLIMB, INDICT, HANDKERCHIEF, IMAGINE, FIFTH, GNAW, RHYME, BUSINESS, MARIJUANA, KNIFE, TALK, MNEMONIC, AUTUMN, JEOPARDY, PNEUMONIA, LACQUER, FORECASTLE, ISLAND, LISTEN, BUILD, WRIST, FAUX PAS, PRAYER, RENDEZVOUS.

The non-uniformity of the letters can be considered a source of redundancy in text and it suggests a way to compress text. Replace each letter with a variable-size code, such that common letters are assigned the shortest codes. Already in the 1840's Samuel Morse understood this principle and used it in his popular code (that some people believe was actually developed by his assistant, Alfred Vail.)

In the study of languages, a digram (or bigram) is a pair of consecutive letters or other symbols. Similar to letters, the distribution of digrams is non-uniform. For example, the most common diagram in English is `th`. As if this isn't enough, we can safely guess that a `th` would be followed by an `e`. These features and others are sources of redundancies in natural languages and they help in devising algorithms to compress meaningful text.

Another source of redundancy in natural languages is the proliferation of vowels, which many consider unnecessary. Consider how easy it is to understand a sentence such as "`if u cn rd ths u cn gt a gd jb w hi pa!`" (from a 1983 poem by James Merrill).

Images are common in current computer applications, and they too can easily be compressed. Raise your head and observe the wall in front of you. As your eye moves from point to point on the wall, the color may gradually vary. At the end of the wall, the last point of the wall and its immediate neighbor, the first point on the adjacent wall, may have very different colors, but in general, points that are near neighbors tend to have identical or similar colors. This feature of images, both natural and artificial, is widely used in many algorithms to compress images quickly and efficiently.

Digital videos, which are also very common, can be compressed by utilizing the fact that scenes in a video tend to vary smoothly, resulting in adjacent video frames that tend to be very similar. A compression method for videos can achieve impressive compression ratios by simply recording the differences between the current frame and its predecessor.

> As for a future life, every man must judge for himself between conflicting vague probabilities.
>
> —Charles Darwin.

Therefore, data, particularly in digital format, can be condensed to some degree no matter the type—whether it be text, pictures, audio, or video—due to redundancies, patterns, and predictability. Simply put, it is not by chance. But to what extent can compression reduce the size? Data is compressible because it is not random, implying that it can be compressed further and further until it seems entirely random. A random data file would hold all the original information, but in a compact form. The technical term would be compressing the data to its entropy. Such unstructured information would be considered pure information, indicating that randomness, often linked to disorder (also known as entropy), is also pure information, a surprising finding.

Algorithmic information theory, the brainchild of Gregory Chaitin, defines information in relation to compression and randomness. The idea is that the information content of a string is equivalent to the length of the most-compressed possible self-contained representation of that string.

A string that consists of n copies of the same symbol x is uniform, has low entropy, and can be maximally compressed to nx. The algorithm needed to compress it is

extremely simple, and since this algorithm measures the amount of information in the string, the string has almost no information. The string 142857142857...142857142857 is simply repetitions of the fraction 1/7 and can therefore be highly compressed by a simple, short algorithm. It also has little information, as measured by the length of the algorithm. On the other hand, a truly random string cannot be compressed by any clever algorithm. It has no redundancies, no patterns, and its individual components are unrelated. None can be predicted by any of the others. The compressed image of such a string must be the original string itself, which is why such a string is considered to have high information content or the maximum entropy.

So much for the relation between information and randomness, but a random string contains no meaning for us. A string of random text is gibberish. An image of random pixels looks to us as noise, as is also a video of random frames. We must therefore conclude that what is meaningful to us is data that has patterns and redundancies, and as a result contains no information. It is the patterns in the data that convey meaning, be it text, images, audio, or anything else. Along the same lines, we can consider a scientific theory an attempt to compress a natural phenomenon that may be familiar to us, into a set of equations that is meaningless to most people, but has precise meaning and can be used to make future predictions.

> Freedom is nothing but a chance to be better.
> —Albert Camus.

◇ **Exercise 7.3:** Sometimes, a gambler might say "I haven't seen 27 come up recently, so it must be due." Can such a claim be true? Can a number be "due" in a random sequence?

◇ **Exercise 7.4:** A gambler might say "I haven't seen 27 come up recently, so it must be cursed." Can such a claim be true? Can a number be "lucky" or "cursed" in a random sequence?

Lest men suspect your tale untrue, Keep probability in view. John Gay.

The *only* rule of chance is that any result is possible. Anonymous.

The first half of life consists of the capacity to enjoy without the chance; the last half consists of the chance without the capacity. Mark Twain.

Never miss a good chance to shut up. Will Rogers.

How dare we speak of the laws of chance? Is not chance the antithesis of all law? Bertrand Russell.

7.4 Benford's Law

In the 1940s, I owned a table of logarithms, similar to the one in Figure 7.3. This thin yellow booklet, chock full of numbers, was my constant companion for several years, helping me with inexact, ballpark calculations for physics problems. After several years of heavy use, I observed that the pages were smudged in a non-uniform manner. The initial pages got the most dirty, whereas the final pages remained fairly clean. In fact, Equation (7.3) shows that the first page was precisely 6.54 times dirtier than the past page (page 9). Those who can still recall those tables might remember that the initial pages had logarithms with a leading digit of 1.

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17

Figure 7.3: An Old Table of Logarithms.

Initially puzzled by this, I soon switched my attention to more important things. It was only years later that I discovered that other, more observant people, had observed it, and had researched it and published unexpected, interesting, as well as practical conclusions. Today, this phenomenon is named after Frank Benford, a physicist with the General Electric Company, who published it in 1938, [Benford 38], but the American astronomer Simon Newcomb had noticed this phenomenon as early as 1883. See Figure 7.4.

Notice that the table of Figure 7.3 has 10 columns which correspond to . This is because the logarithm of a larger number, such as 215, which equals 100×2.15 is the sum $\log 100 + \log 215 = 2 + .3324$

No scientific discovery is named after its original discoverer.

—Stigler's Law of eponymy.

That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones.

—Simon Newcomb, [Newcomb 81], 1881.

It has been observed that the first pages of a table of common logarithms show more wear than do the last pages.

—Frank Benford, [Benford 38], 1938.

The fact that leading digits tend to be small is important to keep in mind; it makes the most obvious techniques of 'average error' estimation for floating-point calculations invalid.

—Donald E. Knuth, [Knuth 69], 1969.

The income tax agencies of several nations and several states, including California, are using detection software based on Benford's Law, as are a score of large companies and accounting businesses.

—*The New York Times*, [Browne 98], 1998.

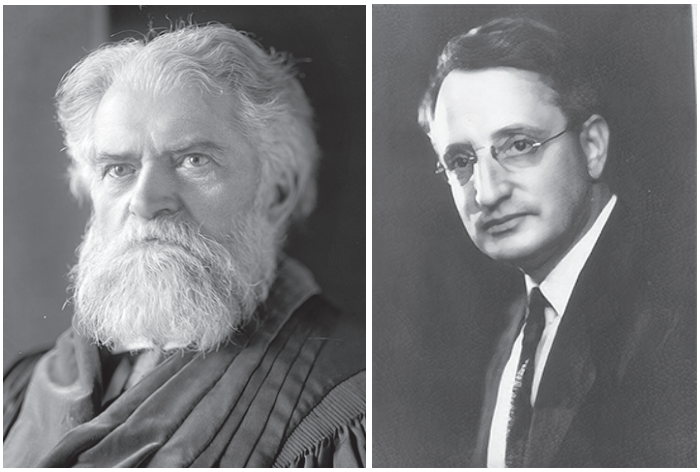


Figure 7.4: Simon Newcomb (left) and Frank Benford (right).

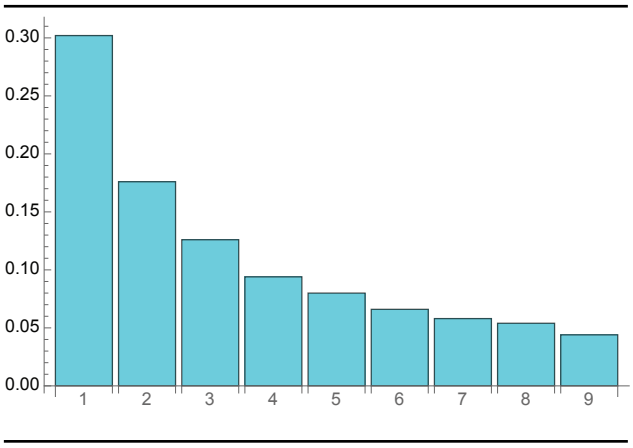
Benford's law, also known as the law of anomalous numbers, or the first-digit law, is a surprising observation. Imagine a large set of real numeric data, data culled from some area of real life. The numbers included in such a set may seem random, they may even be random and may pass all tests of randomness. Specifically, the distribution of digits in the set may be uniform; each of the ten digits may appear about 10% of the time, but there will be at least one hidden feature that would make this data less than random. The distribution of the leading (most-significant) digits of those numbers may not be random and may instead obey Benford's law. Surprisingly, the digit 1 would appear about 30% of the time, 2 would appear about 18%, 3, about 12%, and so on, down to 9, whose prevalence would be only about 5%. (Assuming that zero will never appear as a most-significant digit.) The nine Benford numbers, predicted by Equation (7.4) are

$$0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046. \tag{7.3}$$

Figure 7.5 is a clear illustration of this behavior. This is a bar graph of the distribution of the leftmost digits of the first 500 Fibonacci numbers. It is clear that each successive digit appears fewer times in this sample. Numbers obtained in compound interest calculations also exhibit similar behavior. The elementary Mathematica code `NumberForm[Table[(1 + 0.1)^i, {i, 0, 48}], {5, 2}]` produces the 49 numbers

```
1.00,1.10,1.21,1.33,1.46,1.61,1.77,1.95,2.14,2.36,2.59,2.85,3.14,3.45,3.80,4.18
,4.59,5.05,5.56,6.12,6.73,7.40,8.14,8.95,9.85,10.83,11.92,13.11,14.42,15.86
,17.45,19.19,21.11,23.23,25.55,28.10,30.91,34.00,37.40,41.14,45.26,49.79,54.76
,60.24,66.26,72.89,80.18,88.20,97.02,
```

and it is obvious, even without counting, that many of these numbers start with 1 or 2, and only a few start with 8 or 9.




```
a = Table[0, {n, 9}];
Do[i = First[IntegerDigits[Fibonacci[n]]]; a[[i]]++, {n, 500}]
a
BarChart[a/Total[a], ChartStyle -> {Cyan},
 ChartLabels -> {1, 2, 3, 4, 5, 6, 7, 8, 9}]

[Output]: 151, 88, 63, 47, 40, 33, 29, 27, 22
```

Figure 7.5: Distribution of the Leftmost Digits of the First 500 Fibonacci Numbers.

It is crucial to understand to what sets of data Benford's law applies, so here is the main types:

- Quantitative data and not qualitative data. Quantitative data consists of (continuous) measures or (discrete) counts, and is saved as numbers. Such data describes how much, how many, or how often. Consisting of numbers, a set of quantitative data can be sorted or ordered by size. Numbers of animals in a corral, as also heights and weights of objects in a set, are examples of discrete quantitative data. Examples of continuous quantitative data are sets of pressure values and counts of blood cells.

  Qualitative data consists of descriptions, types, and names that are assigned to quantitative data items. Such data describes an attribute that does not originate from

a measurement. Not being made of numbers, a set of qualitative data cannot be sorted or ordered by size. Examples of such sets are animals' ages, peoples' names, and colors of bolts of cloth.

- The data items in such a set should refer to the dame object, and there should not be a built-in maximum or minimum to the numbers.

- Data that is measured and not simply assigned arbitrarily. Typical examples of assigned data are zip codes, telephone numbers, and sets of identification numbers.

- A set of data must be big enough to provide enough statistics. Also, the data items should range over several orders of magnitudes. A set of numbers that are in the interval $[0, 10]$ may not exhibit the law even if the set itself is large. The more orders of magnitude covered by the data, the more accurately Benford's law applies to it. Examples of important data that do not span orders of magnitude are height, weight, and IQ scores.

- The set of data items must be complete. It should not be artificially restricted, cut, or pruned by the user.

- The data must be real and accurate. If the data is based on units, the law still holds for any choice of units, even though the choice affects the digits, as for example, 12 inches equal 30.48 centimeters.

- Given several data sets where each satisfies Benford's law, the combined set will also obey the law. This is referred to as mixed population. Website `testingbenfordslaw.com` has sources of many data sets that are available for testing the law and its effect on various aspects of life.

The following facts make this law more general and fundamental:

- This behavior of the leading digit holds even when the data is converted between different number bases. The leading digit 1 is the most common one when the same data is stored in decimal form, binary, octal, or any other (integer) number base.

- The law holds for many sets of data taken from widely different sources and at different historical periods. Some examples are street addresses, stock prices, sports statistics, reviews of products, financial data, decision-making documents, tax returns, and billing amounts. In fact, when a large set of real data does not conform to Benford's law, the law may be used as a reason for suspecting the data, although not as a proof of wrongdoing.

Examples of data sets that do not conform to this law are zip codes (in a given region, the leading digits of the codes are too similar), telephone numbers (prefixes may be identical), invoice numbers (may be sequential), birthdays (month and day have built-in limits), prices (most end with one or several 9's), passenger numbers in a list of commercial flights (the limits are set by the size of the plane), and forged documents (the style depends on the author). The square roots and reciprocals of successive natural numbers do not obey Benford's law.

- Later research by [Hill 95] has discovered that not only the leading digit but also the nth leading digit obeys a similar distribution law. Given a data set where the leading digit obeys Benford's law, each following digit obeys a similar logarithmic law which

rapidly approaches a uniform distribution as n gets bigger. Equation (7.6) gives the probability of the first (leading) digit, and the similar rule $P(d_1d_2) = \log(1 + 1/(10d_1 + d_2))$ gives the probability of the pair of first digits d_1d_2. Thus, for example, the probability of 32 as the first two digits is 0.013364, the value of `Log[10, 1.+1/32]`, and the logarithm `Log[10, 1.+1/314]` produces 0.00138091 as the probability of the triplet 314 occurring at the start of a number.

Similarly, the probability of digit d as the second leftmost digit, regardless of the first digit, is the sum

$$\log_{10}\left[1+\frac{1}{10+d}\right]+\log_{10}\left[1+\frac{1}{20+d}\right]+\cdots+\log_{10}\left[1+\frac{1}{90+d}\right].$$

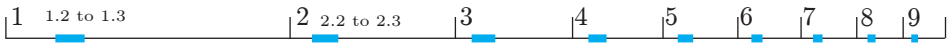
For $d = 2$, this sum is 0.108821, produced by the Mathematica code `Sum[Log[10, 1.+1/(10k+2)], {k,9}]`.

In general, the probability that the digit d appears in position n from the left of a number is the sum

$$\sum_{k=10^{n-2}}^{10^{n-1}-1}\log_{10}\left(1+\frac{1}{10k+d}\right),$$

a sum which both shrinks and becomes less dependent on d for large values of n. At the limit when $n \to \infty$, the sum reduces to Equation (7.4).

The precise probabilities for the second digits being 0 through 9 are 0.11968, 0.11389, 0.10882, 0.10433, 0.10031, 0.09668, 0.09337, 0.09035, 0.08757, and 0.08500, respectively. The figure illustrates how the probability of the second digit being 2 is the total length of the blue segments on the logarithmic scale from 1 to 10. Notice that it is only necessary to measure it on one order of magnitude, because the fraction it occupies is the same on all other orders of magnitude.



The total length is difficult to measure with a ruler, but is easy to compute. It is the fraction

$$\frac{\sum_{k=1}^{9}(\ln(k.3)-\ln(k.2))}{\ln(10)-\ln(1)}\approx 0.108821,$$

or `Sum[Log[k+.3] - Log[k+.2], {k,9}]/Log[10]`

Even more, the joint distribution of the leading n digits also obeys a similar law, which gets weaker as n gets bigger, thereby showing that the n leading digits of the numbers in Benford-conforming data sets are statistically correlated.

As an example, if the leading digit is 3, then the expected Benford probabilities (in percentages) for the second digit being 0 through 9 are 1.42, 1.38, 1.34, 1.30, 1.26, 1.22, 1.19, 1,16, 1,13, and 1.10, respectively.

Table 7.6 lists the distributions of the first (three-significant) digits according to Benford's law. The numbers are in percentages, and it is obvious that already the distribution of the 3rd digit is close to uniform. The difference in the probabilities of zero and 9 is only $10.18 - 9.83 = 0.35\%$.

The following Mathematica code:

```
a1 = Table[Sum[Log[10, 1. + 1/(10 k + d)], {k, 0, 0}], {d, 0, 9}];
a2 = Table[Sum[Log[10, 1. + 1/(10 k + d)], {k, 1, 9}], {d, 0, 9}];
a3 = Table[Sum[Log[10, 1. + 1/(10 k + d)], {k, 10, 99}], {d, 0, 9}];
TableForm[NumberForm[100 {a1, a2, a3}, {6, 2}]]
```

digit:	0	1	2	3	4	5	6	7	8	9
1st:	—	30.10	17.61	12.49	9.69	7.92	6.69	5.80	5.12	4.58
2nd:	11.97	11.39	10.88	10.43	10.03	9.67	9.34	9.04	8.76	8.50
3rd:	10.18	10.14	10.10	10.06	10.02	9.98	9.94	9.90	9.86	9.83

Table 7.6: The Distributions (in Percents) of the 1st, 2nd, and 3rd Digits.

```
a = Table[0, {n, 10}];
Do[i = Last[IntegerDigits[Fibonacci[n]]]; a[[i + 1]]++, {n, 500}]
a
```

counts the distribution of the least-significant digits of the first 500 Fibonacci numbers. It produces the ten numbers $33, 68, 33, 67, 35, 67, 32, 67, 33, 65$. There are twice as many odd digits as there are even digits. This is far from a Benford distribution, and it confirms the little-known fact that the pattern of odd and even numbers in the Fibonacci sequence repeats every three terms: Odd, Odd, Even.

■ Figure 7.7 lists examples of data sets that do not conform to Benford's law. The percentages of the leftmost digits of 500 random integers are listed, and it is clear that they differ significantly from this law. The counts of the leftmost digits of the first 500 positive integers are also listed, and it is obvious that they are completely non-Benford.

```
a = Table[0, {n, 9}];
Do[i = First[IntegerDigits[RandomInteger[{1, 500}]]];
 a[[i]]++, {n, 500}]
a   (*Distribution of the leftmost digits of 500 random integers*)
N[a/Total[a]] (*Percentages*)
a = Table[0, {n, 9}];
Do[i = First[IntegerDigits[n]]; a[[i]]++, {n, 500}]
a   (*Distribution of the leftmost digits of 500 consecutive integers*)

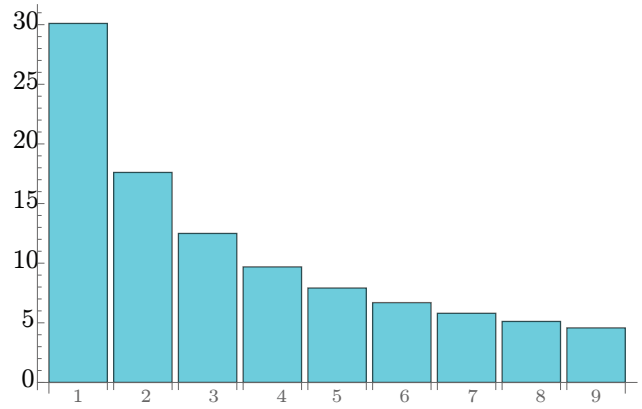
{108, 104, 117, 119, 11, 9, 11, 5, 16} Distribution for 500 random
{0.216, 0.208, 0.234, 0.238, 0.022, 0.018, 0.022, 0.01, 0.032} Percentages of
 leftmost digits of 500 random integers
{111, 111, 111, 111, 12, 11, 11, 11, 11} Distribution for the integers 1-500
```

Figure 7.7: Data Not Exhibiting Benford's Law.

The mathematical form of Benford's law is logarithmic. In a data set that obeys Benford's law, the leading digit d (an integer in the interval $[1, 9]$) occurs with probability

$$P(d) = \log(d+1) - \log(d) = \log\left[1 + \frac{1}{d}\right]. \tag{7.4}$$

Thus, $P(d)$ is proportional to the space between d and $d+1$ on a logarithmic scale, as shown graphically in Figures 7.8 and 7.9b.




```
Table[NumberForm[N[Log[10, n + 1] - Log[10, n]], {4, 3}], {n, 9}]
BarChart[100  Table[Log[10, n + 1] - Log[10, n], {n, 9}],
 ChartStyle -> {Cyan}, ChartLabels -> {1, 2, 3, 4, 5, 6, 7, 8, 9}]

0.301,0.176,0.125,0.097,0.079,0.067,0.058,0.051,0.046
```

Figure 7.8: Probabilities of the Leading Digits.

Explanations of Benford's Law

The following are several attempts to explain Benford's Law. Readers who prefer a single, readable explanation are referred to reference [Fewster 09], a short, clear, well-written explanation with only little statistics.

- Simon Newcomb, the Canadian-American astronomer and mathematician, published in 1881 a "Note on the Frequency of Use of the Different Digits in Natural Numbers." For Newcomb, natural numbers were those occurring "in nature," i.e., the kind of numbers one would run into in the course of everyday life. He discovered, for example, that not all the digits $(1, 2, \ldots, 9)$ occur with the same frequency in the first place of such a number; he formulated a law and gave a rough proof based on the argument that the mantissas (the fractional part) of logarithms are uniformly distributed, but the integer parts are distributed logarithmically.

Definitions of `mantissa`. This is a mid 17th century term meaning "makeweight." Likely from Latin or Etruscan. Its modern use in mathematics is: the part of a logarithm that follows the decimal point. In computing it is used as a significand, the part of a floating-point number that represents the significant digits of that number.

Here is what the terms "integer part" and "mantissa" mean. Recall that $b \stackrel{\text{def}}{=} \log_{10} a$ if $10^b = a$. The following examples $\log_{10} 1.234 = 0.0913$, $\log_{10} 12.34 = 1.0913$, and

$\log_{10} 123.4 = 2.0913$, imply that $10^{0.0913} = 1.234$, $10^{1.0913} = 12.34$, and $10^{2.0913} = 123.4$. Thus, given an arbitrary real number such as 123.4, we can write it as

$$123.4 = 1.234 \times 10^2 = 10^{0.0913} \times 10^2 = 10^m \times 10^n,$$

where m is the mantissa and the 1 is the integer part i (a single nonzero digit to the left of the decimal point). The logarithm of 123.4 is now

$$\log_{10}(123.4) = \log_{10}(10^m \times 10^n) = \log_{10}(10^m) + \log_{10}(10^n) = m + n = n.m, \qquad (7.5)$$

or $2 + 0.0913 = 2.0913$ (the dot in $n.m$ is a decimal point).

Newcomb noticed that in data sets that conform to his observation (today we say, sets that obey Benford's law) the leftmost digits are evenly spread when drawn on a logarithmic scale. This led him to Equation (7.5) and to the claim that the expression

$$P(d) = \frac{\log(d+1) - \log(d)}{\log(10) - \log(1)} = \log(d+1) - \log(d) = \log(1 + 1/d) \qquad (7.6),$$

is the probability of the leftmost digits d. This behavior is illustrated graphically in Figure 7.9a, which shows how a linear plot of the mantissa values from 0 to 1 corresponds to a logarithmic plot of the values of i. Notice that the four intervals 1–2, 2–4, 4–8, and 5–10 in the figure have equal lengths, implying that i has equal probabilities of being in any of them.

The figure also illustrates Equation (7.4), because it shows graphically that, for example, the interval from 4 to 5, which corresponds to real numbers with $i = 4$, equals $\log(5) - \log(4)$. Given a data set of real numbers, the probability of all the data items of the form 123... is therefore $\log(124) - \log(123) \approx 0.0035$. Part (b) of the figure is a large logarithmic scale with the powers of 10 in red and powers of 2 in blue. The distances between consecutive red lines are equal, as also are the distances between consecutive blue dots. This illustrates the relation between logarithmic scale and the Benford distribution.

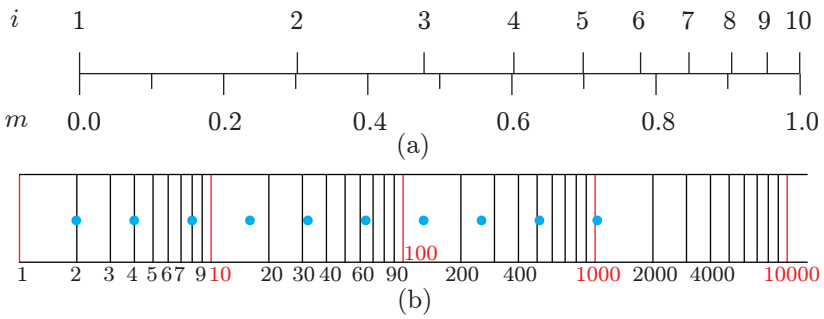


Figure 7.9: Logarithmic and Linear Scales.

- Hill's explanation. This concerns the concept of scale. Consider a data set of integers d_n that are the lengths, in kilometers, of roads, railways, bridges, or rivers in a certain region. Assume that many roads are in the interval 15–30, while only a few long roads are between 150 and 300 kilometers long. We now decide to express all the lengths in hectometers (100 meters), so we multiply them by 10. All the d_n's that were between 15 and 30 are now between 150 and 300 hectometers, while the few d_n's that were between 150 and 300 are now in the interval 1500–3000. We say that the data is not scale invariant; its distribution depends on the units of measure, the scale. The distance between the new intervals 150–300 and 1500–3000 is much greater than the distance between the original intervals 15–30 and 150–300.

  In order for a set of data items to be scale invariant, we examine the logarithms of the two original intervals. They are $\log(15) = 1.17$ and $\log(30) = 1.48$ (for 15–30) and 2.17 and 2.48 (for 150–300). The lengths of the original intervals were 15 and 150, but the lengths of the logarithms of those intervals are both ones; they are equal. We therefore conclude that in order for a set of data items to obey Benford's Law, they have to be scale invariant; their logarithms should be distributed as in the logarithmic scale of Figure 7.9b. Turning this argument on its head we conclude that Benford's Law is a natural consequence of a data set the probabilities of its individual items satisfy Equation (7.4). This is the essence of the proof published by Theodore Hill in 1995, more than a hundred years after Newcomb.

- Here is a slightly different way of looking at scale invariance. Given a data set where the distribution of the first digits is scale-invariant (independent of the units of the data), the data will always obey Benford's law.

  Consider the simple relation of one meter equals three feet (this is approximate, but we'll ignore it). The probability that the first digit of a length in meters is 1—i.e., the length is in the 1-unit interval $[1, 2)$—equals the probability that the first digit of the same length in feet is in the 3-unit interval $[3, 6)$. Similarly, the probability that the first digit of a length in meters is 2—i.e., the length is in the 1-unit interval $[2, 3)$—equals the probability that the first digit of the same length in feet is in the 3-unit interval $[6, 9)$. This is true for all the values of the data items, which implies a logarithmic distribution of Benford's law.

- Following the excellent video at [Barrow 11] (19:30 in the video), it can be shown that data that is scale invariant must obey Benford's law. We can refer to a scale invariant data set as data that has a universal distribution, and we denote the probability of data item x in this set by $P(x)$. The gist of this explanation is that if there is a universal distribution, then it must be Benford's. However, this explanation is lacking because it does not tell us why a universal distribution must exist in the first place.

  The distribution of such data, and in particular its set of probabilities $P(x)$, must be invariant under a multiplication by some scale factor k, such as converting the data units from Euros to Yens. We can therefore write $P(kx) = f(k)P(x)$ for some function $f(k)$ of the scale factor.

  Recall that the sum (or, in the case of a continuous distribution, the integral) of the probabilities of a set of items must be 1, $\int P(x)dx = 1$, which implies $\int P(kx)dx = 1/k$. But we already know that $\int P(kx)dx = f(k)\int P(x)dx = f(k)$, which implies that $f(k) = 1/k$. Differentiating $P(kx)$ with respect to k and setting $k = 1$ gives $xP(x) =$

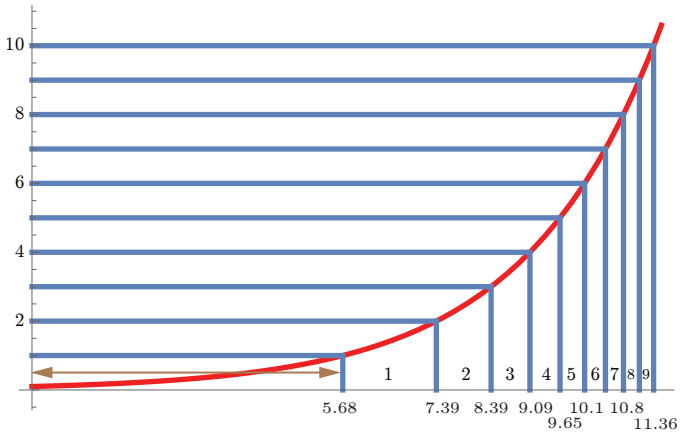
$-P(x)$, a simple differential equation whose solution is $P(x) = 1/x$. (This implies that $\int_0^\infty P(x)dx = \int(1/x)\,dx$ diverges, but in practical applications the data items are never infinite.)

When the data covers many orders of magnitude, the probability that its leading digits d are given by a logarithmic distribution is

$$P(d) = \frac{\int_d^{d+1} P(x)dx}{\int_1^{10} P(x)dx} = \frac{\int_d^{d+1} \frac{1}{x}dx}{\int_1^{10} \frac{1}{x}dx} = \frac{\log(d+1) - \log(d)}{\log(10)} = \log_{10}\left(1 + \frac{1}{d}\right),$$

which is Equation (7.4).

■ An explanation using exponential data. The basic exponential functions are $y = a(1+r)^x$ and $y = e^x$, where a is an initial value and r is the rate of change. Figure 7.10 shows such a curve for $a = 0.1$ and $r = 0.5$. The figure illustrates how the (red) curve rapidly becomes steeper and steeper, and how this affects the gaps between consecutive vertical line segments. The horizontal lines are equally spaced and are placed at integer values of y from 1 to 10. The first vertical segment is placed at $x = 5.68$, where y becomes 1, and the next vertical segment is placed at $x = 7.39$, where y becomes 2.




```
a = 0.1; r = 0.5;
g0 = Plot[a (1 + r)^x, {x, 0, 11.5}, PlotStyle -> {Red}];
g1 = Graphics[ListLinePlot[{{0, 1}, {5.67887, 1}, {5.67887, 0}}]];
g2 = Graphics[ListLinePlot[{{0, 2}, {7.38838, 2}, {7.38838, 0}}]];
g3 = Graphics[ListLinePlot[{{0, 3}, {8.38838, 3}, {8.38838, 0}}]];
g4 = Graphics[ListLinePlot[{{0, 4}, {9.0979, 4}, {9.0979, 0}}]];
g5 = Graphics[ListLinePlot[{{0, 5}, {9.64824, 5}, {9.64824, 0}}]];
g6 = Graphics[ListLinePlot[{{0, 6}, {10.0979, 6}, {10.0979, 0}}]];
g7 = Graphics[ListLinePlot[{{0, 7}, {10.4781, 7}, {10.4781, 0}}]];
g8 = Graphics[ListLinePlot[{{0, 8}, {10.8074, 8}, {10.8074, 0}}]];
g9 = Graphics[ListLinePlot[{{0, 9}, {11.0979, 9}, {11.0979, 0}}]];
g10 = Graphics[ListLinePlot[{{0, 10}, {11.3577, 10}, {11.3577, 0}}]];
Show[g0, g1, g2, g3, g4, g5, g6, g7, g8, g9, g10,
 Ticks -> {None, Automatic}]
```

Figure 7.10: Probabilities of Leading Digits in an Exponential Curve.

The gap between those segments corresponds to all the y values in the interval $[1, 2)$, where the leading digit of y is 1. The size of this gap, $7.39 - 5.68 = 1.71$, satisfies $1.71/5.68 = 0.301$. It is therefore 30.1% of the gap where y is zero (marked with a double arrow line). The corresponding numbers for the next gap, labeled 2, are $8.39 - 7.39 = 1$ and $1/5.68 = 0.176$ or 17.6%. Notice also that the four gaps 1–2, 2–4, 4–8, and 5–10 have the same sizes $7.39 - 5.68 = 1.71$, $9.09 - 7.39 = 1.7$, $10.8 - 9.09 = 1.71$, and $11.36 - 9.65 = 1.71$. This is why between Benford's law is the result of exponential data. Given such data, we can expect 30.1% of the items to start with 1, 17.6% to start with 2, and so on.

Proponents of this explanation point out that many type of common data, such as population growth and financial data, grow exponentially. Exponential data can also decay, and the main features of such Decay are the following:

- Decreasing Rate: Data items shrink faster initially, but the decay rate decreases as the remaining quantity gets smaller.

- Constant Percentage Decrease: The same percentage of the data decreases over equal time intervals.

- Asymptote: The quantity being decayed approaches zero ever so slowly, and may, in practice, be completely gone. In principle, however, it never reaches zero. There will always be some fraction of the initial amount left.

- It is known that any quantity that varies exponentially in time, increasing or decreasing, satisfies Benford's law asymptotically. The percentage of time that it spends with each first digit approaches the theoretical Benford value asymptotically over time.

◇ **Exercise 7.5:** What are real-world examples of data that decays exponentially.

- While these explanations seem reasonable, they do not explain why Benford's law holds for data that is not consecutive integers or exponential. The law is obeyed, among others, by data sets of street addresses, physical constants, and file sizes in a computer. The following statement, from [Hue 13], may provide a possible explanation: If you multiply many random numbers together and repeat this many times, the resulting sequence of random numbers would span many orders of magnitude, as well as obey Benford's law.

Figure 7.11 illustrates an example. Roll a die 20 times, and multiply the 20 numbers which are random integers in the interval $[1, 6]$. The product may be between 1 (all ones) and $6^{20} \approx 3.65 \times 10^{15}$ (all 6's), which is a very wide interval. Repeated many times, the resulting sequence of these random numbers obeys Benford's law, as illustrated by Figure 7.11.

- A related explanation, mentioned in the Wikipedia article on Benford's law, has to do with the fact that the distribution of an exponential data looks linear when plotted on a logarithmic scale. Figure 7.12 shows a narrow distribution, covering the range 10 through 150, and a wide distribution that is spread from zero to beyond 10,000. The main feature is the relation between the widths and areas of the regions that correspond to the leading digits (some of them colored). The red regions are those parts of the distributions where the leading digits are 1, and the blue regions are the parts where

```
a = Table[0, {n, 9}];
Do[{b = 1;
  Do[b = b*RandomInteger[{1, 6}], {20}];
  i = First[IntegerDigits[b]];
  a[[i]]++;}, {1500}]
N[NumberForm[a/Total[a], {5, 2}]]

0.30,0.17,0.11,0.12,0.08,0.08,0.05,0.05,0.05
```

Figure 7.11: Probabilities of Many Random Numbers That Are Multiples.

the leading digit is 8. It is obvious that the regions in the narrow distribution are either narrow and tall or wide and short, whereas the same regions in the wide distribution conform to Benford's law. They are wide for a leading digit of 1 and narrow for a leading digit of 8, everywhere along the range of the data, and regardless of their heights.

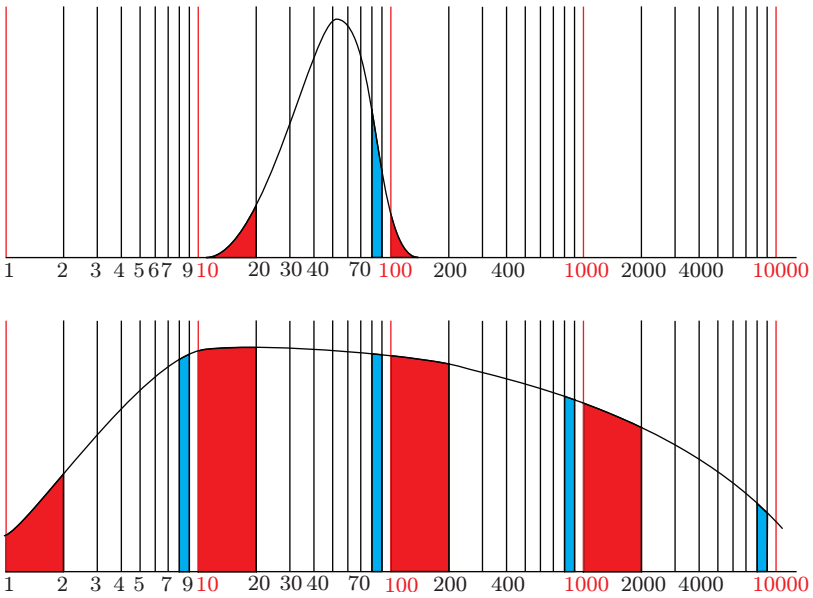
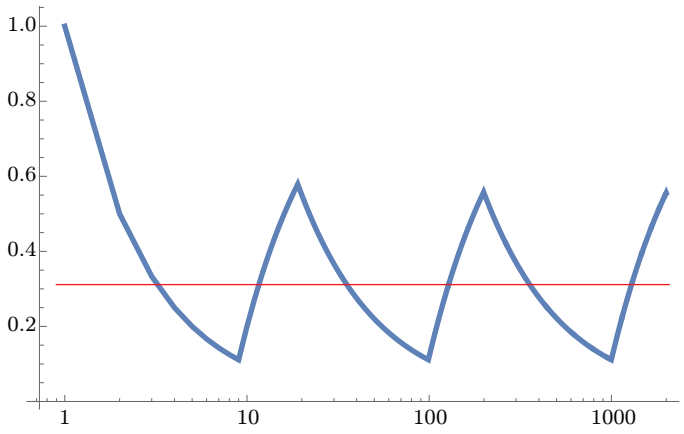


Figure 7.12: Narrow (Top) and Broad (Bottom) Distributions.

This behavior of the regions (colored or otherwise) in the figure, combined with the fact that a wide, flat distribution corresponds to exponential data, reinforces the idea that Benford's law is the result of exponential data.

■ Another explanation is the so-called "raffle tickets." The many tickets sold for a certain raffle are numbered consecutively from 1. Mathematically, the problem is to find the probability of the leading digit being 1 in the first n consecutive positive integers. For $n = 2$, the probability is $1/2$. For $n = 3$ it is $1/3$, and so on until for $n = 9$ it is $1/9$. At $n = 10$ it jumps to $2/10 = 1/5$, then it climbs to $3/11$, $4/12$, and so on up to $11/19$.

Starting at $n = 20$ there are few 1's until we reach 100, where every following integer until 199 contains a 1. The result is the sawtooth curve illustrated in Figure 7.13, where the x-axis is logarithmic.




```
1 a = 0; l = {};
2 (* $a$ counts the numbers of 1's in the integers so far*)
3 Do[i = First[IntegerDigits[n]]; If[i == 1, a++];
4  AppendTo[l, a/n], {n, 2000}]
5 l; (*The probability at each $n$ is appended to list $l$ *)
6 Total[l]/2000.
7 ListLogLinearPlot[l, Joined -> True]
8
9 0.313738
```

Figure 7.13: Probabilities of Digit 1 in Consecutive Integers.

The unusual behavior of this probability of the leading 1's suggests taking the average, and the figure shows that the average probability for $n = 2000$ is 31.37%, very close to the theoretical Benford probability! The interested reader may run the code of Figure 7.13 changing line 3 (`i==1`) to other leading digits to obtain similar results.

This is the "raffle tickets" explanation of Benford's law, and it has its detractors. They claim that the appearance of Benford's law in this case has nothing to do with raffle tickets, but is the result of the data having an exponential distribution. A look at Figure 7.14 shows that about 30% of a logarithmic scale is occupied by numbers that start with 1 (the red parts) and about 17% is occupied by numbers that start with 2 (the blue parts). Therefore, if we sample numbers at random on a logarithmic scale, regardless of any data sets, about 30% of them would start with 1 and about 17% would start with 2. It would be as if you had exponential data.

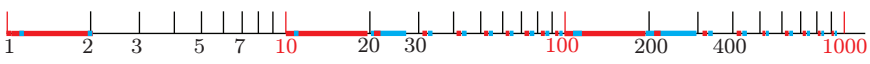


Figure 7.14: A Logarithmic Scale.

Now perform the following tests: (1) Multiply the numbers you just sampled by a fixed factor, the results would still feature the same percentages and would still obey Benford's law. (2) Multiply the sampled numbers by a set of random numbers. The law still holds. (3) Multiply the sampled numbers by the sawtooth data. Same result.

This baffling behavior has nothing to do with a raffle. It is simply because the data items that create the sawtooth curve (variable `l` on line 5 of Figure 7.13) obey an exponential distribution.

Perhaps the most general explanation is that multiplying numbers creates products that get closer and closer to exponentials, and therefore closer and closer to obeying Benford's law. So is this law simply a side effect of multiplication? The following insert is intriguing.

The following summary, based on reference [Trahan 05], is an intriguing attempt to clarify the connection between multiplication/addition operations and Benford's Law.

Alexander C. Trahan, *When Does Randomness End?*, Project #S1223 at the California state science fair 2005.

The objective of this experiment is to determine whether multiplication, exponentiation, or addition will force a set of random numbers [0,1] to conform to Benford's law, with a statistical significance level of 0.05 by Chi-square testing.

Java was used to create random number sets [0,1] (proven not to conform to Benford's Law) then perform operations on this set with additional random number sets.

When multiplying with sets of random numbers, their conformity to Benford's law becomes 0.05 significant after one multiplication and becomes much more significant with more multiplications. When random sets are added, the conformity is not significant. As the number of additions increases, the set becomes less significant. When random sets are used in exponentials their conformity is much better than 0.05 significance.

> Benford's Law, also known as the first-digit law, has long been seen as a tantalizing and mysterious law of nature. Attempts to explain it range from the supernatural to the measure-theoretical, and applications range from fraud detection to computer disk space allocation.
>
> —R. M. Fewster, *A Simple Explanation of Benford's Law*, 2009.

Applications of Benford's law

■ Fraud detection. It has long been realized that Benford's law can serve as a tool to detect fraud in certain types of data sets. This technique has often been used in recent past, especially in cases of tax fraud. Here is how to use the popular Excel spreadsheet software to check a suspicious set of data for compliance with this law. The steps are as follows:

1. Assuming that 5000 data items have been input into column `A`, extract their leftmost digits into column `B` with the `=LEFT(A1:A5000)` command.

2. Prepare the nine constants 1 through 9 in, say, locations `D3` through `D11`.

3. Set location E3 to command COUNTIF with range B:B and criteria D3. Thus, =COUNTIF(B:B,D3), and copy it through to location E11. This will result in the nine counts of leading digits.

4. Set E12 to command =SUM(E3:E11) to sum the nine counts.

5. Set F3 to =E3/E12, and copy this through to F11. These are the Benford percentages (or probabilities) of the leading digits.

6. Set G3 to =LOG10(1+1/E3) and copy through to G11. These are the theoretical Benford probabilities.

7. You can now set location H3 through H11 to the differences between the nine percentages of column E and the correct probabilities of column F.

By examining the nine differences as well as their sum, you'll be able to judge whether they are big enough for you to suspect that the data does not conform to Benford's law. A Bar chart may help.

- Accounting fraud detection. Benford's law can be used to test a document for Benford conformity. The tests may discover irregularities such as many unexpected large numbers in data or a preponderance of numbers that end with 99. The tests may also point to suspicious features in the data, such as quantities of random numbers, many duplicates, and attempts to pad the data with zeros.

At least the following tests should be run: Leading digit test, 2nd digit test, first (most-significant) two digits, first three digits (this may identify digit duplication), and last (least-significant) two digits. Testing the last two digits can identify cases where numbers have been rounded. In real-life commercial data, the last two digits tend to have a uniform distribution with two exceptions: The pair 00 is common, because many quantities may be specified in large, rounded units, and the pair 99 is also common, because of its use in psychological pricing. Thus, this test may be useful in situations where padding numbers or changing them at random is suspected.

The second-order test is sometimes performed to check for data integrity. The data items are sorted, consecutive items are subtracted, and the absolute values of the differences examined. The two leading digits of the absolute values of the differences should exhibit the Benford probabilities listed in Equation (7.3). Any large deviations should be considered suspicious.

The Summation test. The first two digits of a data item can range from 10 to 99, a total of 90 values. The test starts by summing all the items in the entire data set that start with 10, such as 1024, 1.0406, and 10.934. This should be repeated for all the items that start with 11, with 12, and so on up to 99. The assumption is that the resulting 90 sums should feature a uniform distribution. Any sum that is substantially higher or lower than the average is an indication of either a few very large (small) numbers or a large quantity of moderately large (small) numbers in the data. Such an occurrence may raise suspicion and should be investigated further.

The tests should be performed separately on positive and negative numbers. The assumption is that a set of commercial data that is all positive may indicate positive results such as increased income. Attempts to falsify such data would tend to increase the numbers, to move them away from zero, in order to paint a better picture of the business. On the other hand, a data set with negative numbers may indicate losses, and a possible attempt to reduce them by doctoring the numbers and moving them closer

to zero.

- Use in criminal trials. Evidence based on Benford's law is now routinely being admitted in criminal cases.

- Election data. At the time of writing (2025) it is not clear whether Benford's law applies to election data, but researchers and YouTubers have used it to identify or claim irregularities in election results.

- Identifying psychological pricing. The price of an item should depend on the cost of the basic research that led to the item, the costs of manufacturing, distribution, and advertising, and the size of the expected market. It is assumed that the first two or three digits of a price should obey Benford's law. When the distribution of digits in the costs of a series of related items (a line of vacuum cleaners or models of the same car) deviates from Benford's law, it is reasonable to assume psychological pricing (a marketing strategy aiming to create in the consumer a feeling that a listed price is lower than what it really is).

- Lotteries. It is unclear whether Benford's law may help someone increase their chances of winning a lottery, but there may be special cases. Imagine a lottery where each participant chooses a 4-digit number. One number is drawn and all who picked it up share the jackpot. You may not be able to employ Benford's law to help you win, since all 4-digit numbers have equal probability of winning, but it is possible to help you share the prize with only a few others in case you do pick up the winning number.

  The idea is that most people would pick a number based on their personal information, such as a birthdate, address, or telephone number. However, if you believe that many would pick a number based on Benford's law, i.e., which starts with 1, then you can benefit by picking a number that starts with 9, or even 8. This does not affect your chance of winning, but if you do win, you would share the prize with fewer people (some of whom may be readers of this book; just joking).

◇ **Exercise 7.6:** Now that you know all about Benford's law, try to collect free large lists of real data and check each to see whether it conforms to the law, and if not, why.

7.5 Randomness in Dice

A die is a uniform, perfect cube, with its six sides numbered. Children learn very early that they can entertain themselves for a few minutes by repeatedly tossing dice and yelling the numbers obtained in this way. Initially, a child may lift a die, assuming that the number they need is the one on the bottom face of the die, but a guiding adult may explain how choosing the number on the top face is easier, faster, and has the same probability as the number on the bottom face.

It takes a while for the child to grasp the basic concept of probability, a discovery which happens when he realizes that a die has the same chance of landing on any of its six faces. It takes even longer to discover that dice can have other shapes, more interesting than a mere cube, and it is this fact that encourages toy manufacturers to make dice of many shapes and colors, encourages children to visit toy stores and collect

dice, and obliges harassed parents to lecture their children about the evils of wasting money on worthless junk.

Stories abound about probability enthusiasts who tried to prove the fairness of a cube experimentally, by constructing machines that throw a number of dice simultaneously several times a second, photograph the results of every throw, and use computers to analyze the photographs and automatically count the results of more and more dice throws.

However, even the fairest die may not conform to the theoretical results and may not land on each of its six sides precisely one-sixth of the time, because of the following reasons:

- Over many tosses, the die material, often plastic, may wear unevenly, especially at its corners.

- The small dents drilled on the cube's faces for the numbers (see picture on Page 180) introduce a slight, but real, deviation from perfect symmetry. Casinos have been known to correct for this by filling the dents with paint whose specific weight equals that of the die material.

- The behavior of the die at its point of landing depends on the surface it lands on. Landing on a hard surface, on a carpet, on sand, mud, rubber, or on someone's hand may affect the statistically perfect behavior of the die while it was in the air. Specifically, landing on a hard surface may cause the die to tumble a few more times before coming to rest.

It seems that already Archimedes knew that there are only five shapes, known today as the Platonic solids, which are convex and regular as explained below. They are some of the 13 Archimedean solids which are convex polyhedra whose faces are regular polygons and whose vertices are all symmetric to each other.

A Platonic solid is a convex regular polyhedron. One definition of convex is that a line segment that starts on the surface of the polyhedron and travels inside, is completely included in the polyhedron until it comes out of the other side. A regular polyhedron is one that features congruent (identical in shape and size) faces, angles, and edges. Also, the same number of faces meet at each vertex. The five Platonic polyhedra, the Tetrahedron, Cube, Octahedron, Dodecahedron, and Icosahedron are shown in Figure 7.15. For those who like to see these shapes in 3D, reference [wolfram.platonic 24] offers a Wolfram notebook.



Figure 7.15: The Five Platonic Surfaces.

The discussion here is based on reference [Keller et al. 89], where the authors concentrate on the term "fair," define it rigorously, and distinguish between polyhedra that

are fair by symmetry and those that are fair by continuity.

The term "fair by symmetry" is the more important one. A convex polyhedron is fair by symmetry if and only if it is symmetric with respect to all its faces. Any face can be transformed into any other face by means of symmetric transformations (those that transform the polyhedron to itself). Those include rotation, reflection, or any combination thereof. Section 3.2 explains symmetry groups and how such a group consists of all the symmetry transformations of an object. Given a symmetry group where certain group elements (transformations) take any face to any other face, we can refer to the group as transitive on the faces.

All the polyhedra that are fair by symmetry have been identified and classified after many years of research. In addition to the 13 Archimedean solids, the fair-by-symmetry polyhedra include two infinite classes of prisms and anti-prisms. Section 5.4 (Convex Polyhedra with Triangular Faces) of reference [Behnke et al. 74] lists all these solids.

The Rhombic Tricontahehdron

The rhombic triacontahedron, Figure 7.16, got its unusual (ugly?) name because it is the most common 30-sided polyhedron. It is convex and it features 30 rhombic faces, 60 edges, and 32 vertices of two types. The figure makes it easy to see how certain vertices are common to five faces each while only three faces meet at other vertices.

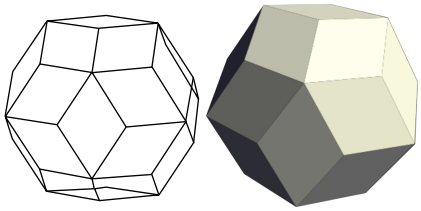


Figure 7.16: The Rhombic Triacontahehdron.

The rhombic triacontahedron is face-transitive; its symmetry group acts transitively on its set of faces. For any two faces A and B, there is a symmetric rotation or reflection that transforms face A to face B. The rhombic triacontahedron is also edge-transitive, and therefore a member of the small, exclusive set of nine edge-transitive convex polyhedra. The other members being the five Platonic solids, the cuboctahedron, the icosidodecahedron, and the rhombic dodecahedron. However, the rhombic triacontahedron is not vertex-transitive. When we hold this solid such that a 5-face vertex is nearest our eyes, and then rotate it such that a 3-face vertex is in the same position, the object would look different! This kind of rotation is therefore not a symmetry.

We thus conclude that the rhombic triacontahedron is fair, but is not as fair as, for example, a cube, because it is not vertex-transitive. This said, it is fair enough so that is has the same probability to land on each of its 30 faces (at least, this is what we currently believe, there being no known experiments to contradict this).

Reference [Rhombictricontahedron 24] shows how to cut and paste a paper model of this intriguing polyhedron.

It has been proved that a die that is fair by symmetry must have an even number of faces, but mathematicians, being both curious and devious, have found a way to go around this result, and the next paragraphs show how fair polyhedra can be constructed (at least in principle) that are fair by continuity but not fair by symmetry.



Consider the figure on the right. It shows a Toblerone chocolate bar, an image that may be familiar to many. For our purposes, this polyhedron is a long prism, so when tossed in the air, it almost always lands on one of its three long, rectangular sides. Now imagine that we cut a thin slice of this prism and also toss it. Naturally, it would tend to land on one of its triangular sides. (Those who are not on a diet tend to dispose of it tastily after the experiment.) Thus, the same object tends to land either on its longer, rectangular sides or on its small, triangular sides depending on its length. Since we can, in principle, cut it to any length, there must be a length where it has the same probability of landing on either type of side. In effect, this unusual solid becomes a five-sided *fair* polyhedron, but in order to distinguish it from fair-by-symmetry polyhedra, we can refer to it as fair by continuity.

> On February 14, 1979, the author [Knuth] bought a box of chocolates and placed the box on a piece of graph paper (after suitably disposing of the contents). The experimental data gathered in this way led to a "definitive" heart shape:
>
> —Donald E. Knuth, *The METAFONTbook*, Addison Wesley, 1986.

Many solid objects, not just polyhedra, can exhibit the same unusual behavior and become fair by continuity. A coin, for example, tends to land on one of its large, circular sides, because its edge is narrow, but if we extrude it and make it look like a sausage, we would end up with a length where the coin, now a short cylinder, has the same probability of landing on its circular ends or on its round body.

A natural question is how long should such a prism or coin be? It is reasonable to guess that the correct length is achieved when the total areas of the two triangular sides equal the total areas of the three rectangular sides, but experiments seem to suggest that the behavior of such a fair-by-continuity polyhedron depends much on the surface it lands on, and factors such as dynamics and surface roughness, elasticity, and suppleness play an important role. This migrates our problem from the realm of mathematics to that of physics.

For the interested reader. Page 95 of reference [Poundstone 95] contains a true anecdote related to such a coin.

The familiar 6-sided cube used as a standard die has many unknown cousins. They are 6-sided polyhedra where the faces are different quadrilaterals with the same area. Such a solid has edges of several lengths, and looks different when viewed from different vertices. However, as claimed in reference [WeirdFairDice 16], its six sides have equal probabilities, which makes it fair, although not fair by symmetry.

◇ **Exercise 7.7:** Explain how to use a standard, six-sided die to draw numbers in the interval $[1, 5]$ with equal probabilities.

◇ **Exercise 7.8:** Can there be a spherical six-sided die?

7.6 Go-First Dice

People love games, and this is especially true for mathematicians. An elementary step in many games consists of a player making a move, and it has been known that in certain games, chess being a notable example, the player who has the first move has an advantage. Reference [first.move 24] proposes several games where first or second movers have statistical advantages.

White has a slight advantage in chess because it always moves first, thereby controlling the direction of the game, and causing Black to always be half a move behind and therefore reacting defensively to White.

> "Who's on First?" is a comedy routine made famous by American comedy duo Abbott and Costello.
>
> —From Wikipedia.

In a game where first or second movers have an advantage, it makes sense to throw dice to decide the order of players. However, ordinary dice create ties when several players roll the same number. An obvious idea that may come to mind is to play several games and change the order of first, second, etc., players. Mathematicians, however, know that $n!$, the number of permutations of n players, grows very quickly with n, so a better idea is to explore the possibility of having a set of special dice, the so-called go-first dice, with the following properties:

- No ties. If a die has a face labeled, say, 4, then no other die has a 4 on its faces.
- Each possible ordering of the players (determined by highest result, next highest result, etc.) has a mathematically equal chance of occurring.
- The above conditions hold for every subset of the whole set. This makes the set useful for games with smaller number of players.

The problem of constructing sets of n go-first dice was researched and solved by several people, most notably Eric Harshbarger [harshbarger 24], starting in 2010.

The first step toward solving any problem is to develop a proper notation, and the notation used here to describe a set of go-first dice has two parts. The first part is a string of the form `2d3+d12+4d6` that indicates two dice with three faces each, one die with 12 faces, and four dice with six faces each.

The second part is an alphabetic string which specifies the labeling of each die's faces. Given a set of n go-first dice, each possibly with a different number of faces, we denote the total number of faces by f and construct a string of size $n \times f$, consisting of n distinct letters, one for each die, where a letter `x` in position k indicates that die x has a face labeled k. Such a string defines the entire set of dice unambiguously, because of the first property above (each label appears once).

Example: The alphabetic string `accbabbcacaa` has an `a` in positions 1, 5, 9, 11, and 12, implying that die `a` has five faces labeled with these numbers. Similarly, die `b` has faces labeled 4, 6, and 7, and die `c` has faces labeled 2, 3, 8, and 10.

◇ **Exercise 7.9:** Given an alphabetic string that is also a palindrome (has mirror symmetry), what can we conclude about the arrangement of the faces of the dice?

The second property above implies that a set of go-first dice must be fair, so the concept of fairness must first be defined rigorously, and it turns out that there are various levels of fairness, three of which are described here in increasing levels of difficulty.

- Go-First fairness. This is fundamental to the concept of go-first and it is obvious. When rolling go-first dice with this fairness, each player has the same chance of rolling the highest number.

- Place fairness. This is the case where, in addition, each player has the same chance of rolling the second highest number, the third highest, etc.

- Permutation fairness. In addition to the above conditions, every possible ordering (permutation) of players must have an equal chance of occurring. (This is unclear to me.)

When constructing a set of go-first dice, it is also possible to consider any symmetries that they exhibit. The following three types are basic symmetries:

- Column Grouped symmetry. Arrange the dice of the set in rows and write the numeric labels of each die in ascending order. The set of three dice of Table 7.17(left) exhibits column grouped symmetry because the smallest values, 1, 2, and 3 are distributed in the three dice, the next smallest values, 4, 5, and 6 are also distributed in the three dice, and so on.

1	6	8	12	13	17		1	8	11	14	19	22	27	30	35	38	41	48
2	4	9	11	15	16		2	7	10	15	18	23	26	31	34	39	42	47
3	5	7	10	14	18		3	6	12	13	17	24	25	32	36	37	43	46
							4	5	9	16	20	21	28	29	33	40	44	45

Table 7.17: Symmetries in Dice Sets.

- Column Paired symmetry. Table 7.17(right) shows a set of four dice, the faces of each are listed in ascending order. Notice that the first two columns of each row of the table add up to 9, the next two columns add up to 25, and so on. It is not clear what advantage this type of symmetry bestows on the set of dice.

- Mirror Symmetry. This is described in Exercise 7.9.

◇ **Exercise 7.10:** So far, this has been pretty complicated, but mathematicians enjoy complexity. However, before reading ahead, I want you, the reader, to explain why this go-first business is actually a solution calling for a problem, and why there are simpler ways to solve the problem of who is first.

Finally, some results, described in [harshbarger 24] with names and dates of discovery.

For three players, a set 3d6 defined by `acbbacbccaabacbbac`.

For four players, a set 4d12 defined by

abcddcbadbaccabdcbaddabccbaddabcdbaccabdabcddcba.

For five players, the impractical set d24+d32+d48+d72+d90 defined by

aaaaaaabcdedccccdedcbbbcdedccccdedcbbbcdedccccdedcbaaaaaaaaaaaaaaaaaaaa.
aaaaaaaaaaaabcdedccccdedcbbbcdedccccdedcbbbcdedccccdedcbaaaaaaaaaaaabcd
edccccdedcbbbcdedccccdedcbbbcdedccccdedcbaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aabcdedccccdedcbbbcdedccccdedcbbbcdedccccdedcbaaaaaaa.

7.7 Subjective Probability

The term "probability" is an indicator of how certain or uncertain we are that an event will occur. This (un)certainty is based on rigorous analysis, on recorded observations, on a long history of collected data, and on precise computations. Thus, a better, more precise term for probability is "objective probability," but this term is too long for practical use. In contrast to objective probability there is also the important area of subjective probability.

Subjective probability is any case where we apply our opinions and beliefs (gut instinct or personal judgment), rather than rigorous analysis or formal calculations, to determine or estimate probabilities. Here are a few typical examples

- It snowed yesterday and it is snowing today. I have every reason to believe that it will snow tomorrow. (Even though we know that the weather can change dramatically overnight.)

- My girlfriend was very late yesterday as well as the day before. Chances are she will be at least 30 minutes late today. I am on time, but why wait for her outside in this rain?

- I am late for my dentist appointment, but my main problem is her forthcoming bill. I doubt if it would be less than 2000 dinars.

In contrast with objective probability, which produces a precise number, subjective probability is normally flexible and is expressed in terms of an interval (believe me, the chances are no more than 20–30%).

One advantage of subjective probabilities is that they keep their owners happy. Thinking that he knows the chance of a future event, helps keep a person satisfied. Another advantage is the intuitive knowledge provided by a guesstimate. A guesstimate is based on previous knowledge, and is therefore a form of subjective probability.

On the other hand, it is easy to see the downside of subjective probabilities. Its main negative features are: (1) A subjective probability reflects the personality and belief of someone, and may therefore be wildly inaccurate. (2) The subjective probabilities of two persons of the same event are the results of high degrees of personal bias. They may therefore be very different and may clash and cause serious disagreements and other negative effects. (3) A person who comes up with a subjective probability may know nothing about probabilities. Specifically, he may not realize that if the probability of an event is p, then the probability of its opposite is $1 - p$.

◇ **Exercise 7.11:** Come up with both good and bad examples of common subjective probability.

> Our brains are just not wired to do probability problems very well.
> —Persi Diaconis.

7.8 Probability and Psychology

So far we have considered rational behavior in our discussion of probability, but psychology and other aspects of human behavior also play a part in the study and understanding of probability. One aspect of the effect of psychology in probabilistic processes is the difference between risk and uncertainty.

Imagine a test that consists of choosing one among several alternatives whose probabilities are known. The right choice results in a large benefit to the subject being tested. We may say that this kind of choice involves risk. Testing many subjects and recording their responses may teach the examiner something about human nature and its approach to risk taking. Now consider a similar test, but this time the probabilities of the options are unknown. This time, the choice involves uncertainty (or ambiguity) rather than risk. Testing a large number of subjects may also lead to a greater understanding of our behavior under conditions of uncertainty. Now imagine two such tests that are similar, so that their results can be directly compared. The following paradox has been proposed by Daniel Ellsberg:

In step 1, an urn contains 100 red and black balls with unknown proportion. The subject being tested is given a choice of two options, to get a prize if a red ball is drawn from the urn or the same prize if a black ball is drawn. This test involves uncertainty, and experiments on a large number of participants have shown that most subjects choose an alternative at random.

In step 2, an urn contains 50 red and 50 black balls, so the probability is known and is also obvious. The subject is given similar alternatives, get a prize if the next ball drawn is red or get the same prize if it is black. There is risk in this alternative, but no uncertainty. In extensive experiments, it was again discovered that most participants choose an alternative at random, even though the probabilities are known.

The final test proposes the following three alternatives to the subject: A ball will be drawn from one of the two urns. You get a prize if it is red. Do you prefer the ball to be drawn from urn 1, from urn 2, or do you have no preference? The results of many experiments indicate that most participants prefer urn 2. One possible conclusion that can be drawn from this is that we feel more comfortable dealing with events and options whose probabilities are known.

7.8.1 The Psychology of PIN Choosing

A picture is worth a thousand words, and the following text demonstrates one consequence of this adage. Scientists, statisticians, and public opinion researchers know that it is often possible to learn much about a subject if data on it is graphed in a certain way. Presently, in 2024, the average person is likely to be the owner of many passwords, identity numbers, and pins (short for Personal Identification Numbers).

Many pins—including those that protect our bank accounts, smart phones, and other digital devices—are only four digits long, which provides for easy memorization and entry. Such pins should be random, but are they? It is reasonable to assume that human nature, being what it is, would cause many users to prefer certain pins, such as symmetric numbers, easy-to-remember numbers, or pins that are based on dates of birth, telephone numbers, national identity numbers, and similar sources.

Reference [youtube.twit 24] presents an analysis of about 3.4 million 4-digit pins in the form of a 100×100 chart consisting of 10,000 dots of various shades of orange. Dots with bright shades correspond to popular pins, while black dots indicate pins that no one had chosen. The data was obtained from several data breaches. The chart demonstrates how easy it is to come up with useful and enlightening information about how people choose such pins. The chart itself is not shown here for copyright reasons, but can be clearly seen in the reference youtube video. Here are some interesting (and disappointing) conclusions about peoples' choices of pins that can be drawn from the chart.

- The bright diagonal line indicates that many pins are of the form $xyxy$. Some of these, such as 0000, 6666, 6969, 7777, and 9999 are very popular. It is now obvious that a hacker trying to guess a pin, would try all 100 possibilities of these numbers, starting with the 10 ones of the form $xxxx$.

- Numbers with numerical runs (consecutive digits), such as 1234, 4321, 4567, 5678, 6789, and 0987 are popular.

- Birth years from 1930 to 1999 and from 2000 to 2004 are common choices. Those are people in the obvious age groups to be current computer users.

- The bottom-left part of the chart, the area from 0000 to 2530, is brighter than the rest, indicating that birthdates, month (1 to 12) and day (1 to 31), are preferred choices for pins. Rows 10–12 are especially bright, which indicates that many people choose a birthday of the form `mmdd` as opposed to `ddmm`. Choosing a birthday to protect important information is obviously risky.

- Numbers taken from literature or movies also tend to be popular. Examples are 2001, 1984, 0007, and 0070. However, numbers drawn from science and mathematics, such as 3141 (π), 2718 (e), and 1618 (ϕ), are rare, indicating that most pin users are non-scientifically savvy.

- When the original 100×100 color chart is converted to grayscale (Figure 7.18), another unusual pattern emerges. It becomes obvious that there are 100 distinct 10×10 grids of cells, each with shades of gray varying uniformly between opposite corners. This reflects the tendency of people to choose pairs of nearby digits, such as 23, 89, and 67, while avoiding distant pairs such as 29, 71, and 38.

- There are about 40 cells that are black or almost black, indicating that only a few users (or none) have chosen them. They are mostly in the top one-third of the chart. See, for example, 6806, 6827, and 6835 (black) and 7036, 7065, and 9938 (almost black). Cells 8437 and 8438 are adjacent and similar, one black and the other almost so. The pair 8397 and 8398 is a similar unexplained example of adjacent and similar cells. An extreme example of opposites is the pair 8146 (black) and 8246 (very bright).

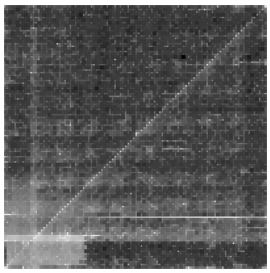


Figure 7.18: 100 10×10 Grayscale Grids of Pin Numbers.

What is the probability that a 4-digit pin will not be chosen by any of 3.4 million users? There are 10,000 4-digit numbers, so the probability of any one being chosen at random is 10^{-4}, and the probability of it not being chosen is $(1 - 10^{-4}) = 0.9999$. The probability that none of our 3.4 million users will not choose this pin is therefore $(1 - 10^{-4})^{3.4 \cdot 10^6} \approx 2.15 \cdot 10^{-148}$. This is so incredibly small, that the fact that there are about 40 such pins implies that pin choosing is not random but is affected by psychological factors.

■ Overall, it is intuitively clear that the distribution of pins is far from uniform, and the following fact makes this obvious. The 20 most common pins—0000, 1004, 1010, 1111, 1122, 1212, 1234, 1313, 1444, 2000, 2001, 2222, 3333, 4321, 5555, 6666, 6969, 7777, 8888, and 9999—account for 27% of the 3.4 million pins that make up the chart. At the opposite side of the distribution, the 21 least-common pins—8557, 9047, 8438, 0439, 9539, 8196, 7063, 6093, 6827, 7394, 0859, 8957, 9480, 6793, 8398, 0738, 7637, 6835, 9629, 8093, and 8068—account for only 2.2% of the 3.4 million pins. Simple, convincing statistics, but don't use 8068 because of an obvious reason.

The conclusion is that a user should consider consider choosing either (1) one of the black cells in the chart, or (2) a 4-digit number that is meaningless to the user, or (3) a longer pin, preferably one that also includes letters and symbols.

See also [datagenetics 24].

◇ **Exercise 7.12:** Why would pin 2580 be common?

7.9 The Birthday Paradox

Given a group of $n > 1$ people, what is the chance that at least two of them will have the same birthday (month and day, not year). Obviously, the chance depends on n, and the surprise is that even with only 23 people, the chance already exceeds 50%. Thus, this problem (see also Page 16) is not a paradox (it does not involve any logical contradiction), but is a surprise, one of many unexpected, unintuitive results that involve probabilities.

To compute this probability, we ignore leap years and also assume that birthdays are evenly distributed throughout the year. (This is not completely true. For example,

if hospitals do not schedule Caesarian sections on weekends, then more babies are born on Mondays and Tuesdays than on weekends. It is also known that more children than average are conceived during long holiday periods, such as Christmas.)

We define event E as: At least two of n people will share the same birthday. The opposite event $\bar{E}$ is therefore: None of the n people will share the same birthday (or alternatively, each of them will have a different birthday). There are 365 days in a year, so at most 365 people can have distinct birthdays. A 366th person will always share a birthday with one of the original 365. If we denote the probability of event E by $P(E)$, then $P(E) + P(\bar{E}) = 1$ and $P(\overline{366}) = 0$, implying $P(366) = 1$.

We start by computing $P(\bar{E})$ for various values of n. For $n = 2$, the first person could be born on any of the 365 days. In order for the second person not to share a birthday with the first one, he must be born in one of the remaining 364 days. The chance of this happening is 364/365. This leaves 363 days for a third person. Thus, the chance that a third person does not share a birthday with either of the first two is 363/365 (we assume that births are independent events, which allows us to multiply their probabilities), and the probability that all three will have distinct birthdays is therefore (364/365)(363/365) or about 0.9918.

Given four people, the chance that they will not share birthdays is, similarly

$$\frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \approx 0.9836,$$

and in general, given n people, the probability is

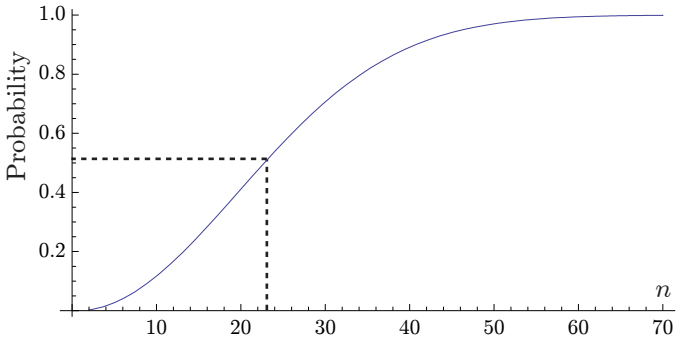
$$\frac{(365-1)}{365}\,\frac{(365-2)}{365}\,\frac{(365-3)}{365}\,\frac{(365-n+1)}{365} = \frac{365!}{(365-n)!\,365^n}.$$

The rest is easy. We know $P(\bar{E})$ and can compute $P(E)$ as $1 - P(\bar{E})$ for any value of n. Figure 7.19 shows the results graphically for $n = 2$ through 70 (for greater values of b, the probabilities are very close to 1). The figure also illustrates the fact that for $n = 23$, this probability is slightly greater than 50%.

■ Finding a birthmate. What is the smallest number of people you need to ask in order to have a 50% probability of finding someone who shares your birthday (month and day)?

This is similar to the birthday paradox with the difference that once a candidate is found, the process stops. The technique used here is to first find the opposite probability, the probability that the next n persons questioned would NOT share your birthdate. Ignoring leap years, we start with $(N-1)/N$, the probability that the next person would "miss" your birthdate (where $N = 365$ is the number of days in a year). Therefore, the probability that the next n persons would have birthdates different from yours is $[(N-1)/N]^n$, and the probability that after questioning n people you would find what you are looking for, is the inverse

$$1 - \left(\frac{N-1}{N}\right)^n = 1 - \left(\frac{364}{365}\right)^n \approx 1 - 0.997^n.$$




```
a = 365!;
Plot[1-a/((365-n)! 365^n), {n,2,70}, AxesOrigin->{0,0},
AspectRatio->.5,
TextStyle ->{FontFamily->"CMR10", FontSize->10}]
```

Figure 7.19: Probabilities of Sharing a Birthday Among n People.

Thus, the answer is the solution of the equation $1 - 0.997^n = 0.5$ or $0.997^n = 0.5$, which is solved by $n \log(0.997) = \log(0.5)$, producing $n \approx 230.7$. The surprising answer is, slightly more than 230 persons may have to be interrogated.

7.10 Choosing a Candidate

In the 1964 movie *Mary Poppins*, Mr. George W. Banks decides to choose a new nanny for his children Mike and Jane. He advertises in the *The Times*, and pretty soon there is a line of candidates at his door, 17 Cherry Tree Lane. He expects them to come in one by one and be interviewed, but instead he unwittingly chooses the very first one, the enigmatic Mary Poppins, thereby providing us with a delightful 138 minutes of pure entertainment.

Now for real life (see also reference [Symonds 14]). You have a vacancy in your organization and you want to hire the best person for the job. You advertise it and a total of N people apply. This is too many for you to interview, so you decide to use the power of mathematics, specifically probability theory, for an optimal way to choose one. He may not be THE best, but he may be the best that can be chosen after interviewing only K of the N applicants.

Your plan is to interview only K applicants, find the best of them, say person B, and then continue to interview applicants until you find the first who is better than B. This applicant is then chosen and the remaining applicants are politely told that the position had been filled.

The problem is to determine the best value of K as a percentage of N. Naturally, this solution is a compromise. We may never interview the BEST applicant, but on the other hand, we may save much time, especially if the optimal K turns out to be much smaller than N. For the mathematical treatment, we have to distinguish between the first K applicants and the remaining $N - K$ ones.

Y-yes, sir?
Tell the other applicants they may go! The position has been filled!

—From *Mary Poppins*, 1964.



Figure 7.20: Image by MS Copilot, Sept. 2025.

The probability of any of the N applicants waiting in line in position i is $1/N$. For each applicant in the first K positions, the probability of him being chosen is zero, while for the other applicants, from position $K+1$ to position N, the probability of anyone being chosen is slightly more complex.

Let's examine the applicant in position $K+2$ and think negatively. If this person is NOT chosen, it is because one of his $K+1$ predecessors had been chosen. The probabilities of any the first K applicants to be chosen are all zeros, and the probability of the $(K+1)$st is $1/(K+1)$. Therefore, the probability of the $(K+2)$nd person to BE chosen is $1-1/(K+1) = K/(K+1)$.

Once this is grasped, the probability of success is given by the sum

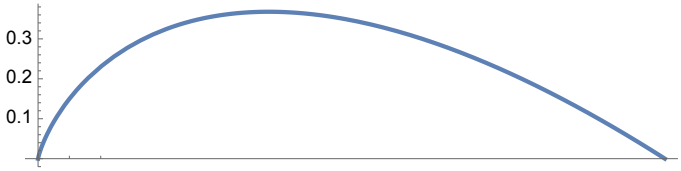
$$\begin{aligned} P(K) &= \sum_{i=1}^{K} \mathrm{P(applicant\ being\ in\ position\ i)P(applicant\ i\ is\ chosen)} \\ &= \sum_{i=1}^{K} \frac{1}{N} \times 0 + \sum_{i=K+1}^{N} \frac{1}{N} \frac{K}{i-1} \\ &= \frac{K}{N}\left[\frac{1}{K} + \frac{1}{K+1} + \cdots + \frac{1}{N-1}\right]. \end{aligned} \tag{7.7}$$

We now notice that the sum in the square brackets is a discrete, approximate version of the continuous function $1/x$ from $K+1$ to N (where x takes the place of K/N). The definite integral for $1/x$ is

$$\int_{K+1}^{N} \frac{1}{x} dx = \ln(x)\Big|_{K+1}^{N} = \ln(N) - \ln(K+1) = \ln\left(\frac{N}{K+1}\right),$$

Which allows us to approximate Equation (7.7) by substituting $x \approx K/N$, to get $P(x) \approx x\ln(1/x) = -x\ln(x)$, a simple function whose maximum is easy to compute by equating its derivative to zero. Thus, $P'(x) = -\ln(x) = x(1/x) = -\ln(x) - 1$, and equating this to zero produces $x = e^{-1} \approx 0.368\ldots$.

Since the continuous variable x takes (approximately) the place of the discrete K/N, we conclude that the optimal value for K is $0.368N$ or about 37% of N. Thus our successful strategy is to interview, rank, and reject the first 37% of applicants and choose the first one after that who is better than the best rejectee. Figure 7.21 lists Mathematica code to plot $P(x)$ and compute its maximum.




```
f[x_]:=-x Log[x];
Plot[f[x], {x,0,1}, AspectRatio->.25]
FindMaximum[f[x],x]
```

Figure 7.21: Function $-x\ln(x)$ and its Maximum.

> Ignorance gives one a large range of probabilities.
> —George Eliot, *Daniel Deronda II.*

◇ **Exercise 7.13:** An employer needs to hire a smart assistant. He advertises, interviews applicants, and selects three seemingly intelligent candidates, A, B, and C. He then passes them through the following intelligence test: "I am going to turn off the light and draw crosses on your foreheads in white or black chalk. When the light is finally turned on, look at the foreheads of the other two candidates. If you see a black cross, raise your hand. If you can tell the color of the cross on your own forehead, say 'stop'."

The employer then draws three black crosses in complete darkness and turns the light on. All three immediately raise their hands, and after only a few seconds candidate C says "stop. I have black."

How did C know his cross color?

7.11 Examples of Unexpected Probabilities

The following examples are mostly drawn from reference [Bogomolny 20]), and their solutions were contributed by many volunteers. They were included here because they are easy to understand and their solutions are short and mostly unexpected.

■ The 1965 movie *Paris Vu Par* (Paris seen by) is a set of six episodes, one of which is titled *Montparnasse et Levallois*. Directed by Jean Luc Godard, this vignette describes the tribulations of a young woman in Paris. She writes and places letters to her two lovers in envelopes, addresses and seals them, and drops them in the pneumatiques. Later, she starts suspecting that she had mixed up the letters (she did not). This vignette gave me the idea for the following cute problem.

A busy secretary writes four different business letters and places them in four envelopes addressed to different recipients. What is the probability that exactly three envelopes will have the correct letters?

Answer. Once three letters go in the correct envelopes, the fourth letter has only its correct envelope left and must go into it. The probability of it going in a wrong envelope is accordingly zero.

■ In 1962, on the occasion of his receiving a second Nobel prize, in peace, the well-known chemist Linus Pauling claimed, perhaps as a joke, that the probability of receiving this prize a second time is actually large, because only a few hundred living people have ever received this prize once. Thus, while it is remarkable to receive a Nobel once, it is much less so to receive it again. What is the fallacy of this argument?

Answer. Receiving a Nobel prize, first or second, has nothing to do with the number of those who had received the prize in the past. The prize committee decides on the winner regardless of whether or not he has already won the prize. Thus, the chance of winning a second Nobel (or any similar) prize is the same as winning a first one.

> I picked up a hitchhiker the other day in my car. He said, "how do you know I am not a serial killer?" My response, "the chance of having two serial killers in the same car is astronomically small."
>
> —Anonymous.

■ Given a 7×7 matrix, fill it up with the digits 1 through 7 such that each column will have all seven digits and the matrix will also be symmetric about the main diagonal. What is the probability that the diagonal has all the integers from 1 to 7?

Answer. The point of this problem is that the matrix is symmetric and also has an odd number of columns. We concentrate on occurrences of the digit 1. This digit must appear seven times in the matrix and must appear exactly once on each column. For each occurrence of this digit off the diagonal, there is another, off-diagonal symmetric occurrence of it on a symmetric column, for a total of six occurrences. Thus, the seventh occurrence of the digit 1 must be on the diagonal. This argument is true for each of the seven digits, which means that each appears on the diagonal and the probability in question is 1.

■ We learn from Wikipedia that a loaded, weighted, cheat, or crooked die is one that has been tampered with so that it will land with a specific side facing upwards more often (or less often) than a fair die would. There are several methods for making loaded dice, including rounded faces, off-square faces, and weights.

A group of organized crooks may load a bunch of dice identically in order to cheat wholesale in a set of games. Your task is to show them that when playing with identically-loaded dice, the probability of getting a double is always at least 1/6.

Answer. This is easy. We denote by p_k the probability of such a loaded die to end up with side k on top. The sum of the six p_k's is 1, which enables us to write

$$\begin{aligned} 1 = 1^2 &= (p_1 + p_2 + p_3 + p_4 + p_5 + p_6)^2 \\ &= (1 \cdot p_1 + 1 \cdot p_2 + 1 \cdot p_3 + 1 \cdot p_4 + 1 \cdot p_5 + 1 \cdot p_6)^2 \\ &\le (1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2)(p_1^2 + p_2^2 + p_3^2 + p_4^2 + p_5^2 + p_6^2) \\ &= 6(p_1^2 + p_2^2 + p_3^2 + p_4^2 + p_5^2 + p_6^2), \end{aligned}$$

Or $(p_1^2 + p_2^2 + p_3^2 + p_4^2 + p_5^2 + p_6^2) \ge 1/6$. (The inequality above is referred to as Cauchy–Schwarz inequality.)

■ According to Wikipedia, the George Washington Bridge, which opened in New York in 1931, is the world's busiest vehicular bridge. The bridge carries 14 lanes of traffic, seven in each direction. This is in addition to pedestrians and cyclists, who use its upper deck, often getting soaked by rain.

The bridge in our problem is much smaller, taking just five minutes to cross by foot, and is also much less used, with only 1000 people cross it during the 12 hours it is open daily. To simplify your task, we assume that the density of crossers is uniform and does not depend on the time of day. Under these assumptions, find the probability that the bridge will be empty at noon.

Answer. An empty bridge at noon implies that no one has entered it since 11:55 AM. There are 144 5-min intervals in 12 hours, which is why, given such an interval, the probability that a person will enter the bridge during that interval is 1/144. Therefore, the probability that none of the 1000 daily people using the bridge will enter it during a given interval is

$$\left[1 - \frac{1}{144}\right]^{1000} = \left[\left(1 - \frac{1}{144}\right)^{144}\right]^{\frac{1000}{144}} \approx e^{-\frac{1000}{144}} \approx 0.000963976.$$

This small probability applies to any 5-min interval, not just at noon.

> Nature permits us to calculate only probabilities.
> —Richard Feynman, *QED: The Strange Theory of Light and Matter*, 1985.

■ With n an even positive integer, choose n real numbers a_1 through a_n at random. Consider them points on the real number line and examine the $n/2$ segments defined by the pairs $a_1\,a_2$, $a_2\,a_3$, $\ldots$, $a_{n-1}\,a_n$. (Notice that sometimes a_i may be greater than a_{i+1}.)

What is the probability that one of these segments would be principal (i.e., include all the other ones)?

Answer. This is easy. We are looking for an index k such that a_k would be the smallest of the n numbers and a_{k+1} would be the largest. Suppose that a_5 is the smallest random number chosen. It is paired with a_6, and the probability of a_6 to be the largest of the $n-1$ remaining random numbers is $1/(n-1)$. A quick check is the case $n=2$, where there are two numbers and one segment, whose probability is $1/(2-1)=1$.

Another approach to the answer is more complex: The n indexes can be permuted in $n!$ ways. In order for a principal segment to exist, there must be two extreme indexes k and $k+1$. The remaining $n-2$ indexes can be permuted between these two extremes in $(n-2)!$ ways. Thus, for a principal segment to exist, only $m=2(n-2)!$ permutations of the total of $n!$ should be considered. For each of the m permutations, a different segment is principal, which is why the probability of there being a principal segment is

$$\frac{n\cdot 2(n-2)!}{n!}=\frac{1}{n-1}.$$

◇ **Exercise 7.14:** (Half jokingly). What is the chance that three people and a monkey who just happen to be in the same location would all be left handed?

■ You are given n cards with the integers 1 through n written on them, one per card. Shuffle the cards and draw m cards. What is the probability that they are drawn in their natural order, from smallest to largest?

Answer. The m drawn cards can be in any of $m!$ permutations, but in only one of which they are in their natural order. Thus, the probability that they have been drawn in that permutation is $1/m!$. Simple.

> A philosopher once said "It is necessary for the very existence of science that the same conditions always produce the same results." Well, they do not.
>
> —Richard Feynman, *The Character of Physical Law.*

■ The outside faces of an $n\times n\times n$ cube (for $n>1$) are painted red. The cube is cut into n^3 unit cubes which are mixed up and placed in a bag. One small cube is withdrawn at random from the bag and tossed across a table. What is the probability that the cube stops with a red face on top?

(The English phrase "Our faces are red" means we are embarrassed.)

Answer. The original cube has six faces, each of size $n\times n$. The total number of small red faces is therefore $6n^2$. Once it is cut, each of the n^3 small cubes has its own six faces, for a total of $6n^3$ faces. As a first approximation we can say that the probability of picking a red face is simply $6n^2/(6n^3)=1/n$, but this has to be further checked and refined, because some cubes have up to three red faces while others, which used to be located inside the original cube, have no red faces. So let's count the small cubes in groups according to how many red faces each has.

Group 1. Each of the eight small cubes at the corners of the large cube has three red faces.

Group 2. Each of the six faces of the large cube consists of four columns of $n-2$ small cubes each. The total is $6 \times 4 = 24$ columns, but this counts each column twice. Thus, this group consists of $12(n-2)$ small cubes, each of which has two red faces.

Group 3. Without its corners and surrounding columns, each large face has $(n-2)^2$ small cubes, each with only one red face. The total number of this group is $6(n-2)^2$.

Group 4. The interior small cubes. We simply peel off the outside faces of the large cube, to end up with an $(n-2)\times(n-2)\times(n-2)$ cube, having $(n-2)^3$ small cubes, each without any red faces.

The total number of small cubes is therefore $8 + 12(n-2) + 6(n-2)^2 + (n-2)^3$, and this really equals n^3. Now for the probabilities. We first define a function $f(k,n)$ that returns the number of small cubes with k red faces

$$f(k,n) = \begin{cases} (n-2)^3, & \text{for } k=0; \\ 6(n-2)^2, & \text{for } k=1; \\ 12(n-2), & \text{for } k=2; \\ 8, & \text{for } k=3; \\ 0, & \text{for } k=4,5,6. \end{cases} .$$

With this function, the probability we are looking for is straightforward

$$\begin{aligned} P(n) &= \sum_{k=0}^{6} P(\text{red face up}|\text{cube has } k \text{ red faces}) P(\text{cube has } k \text{ red faces}), \\ &= \sum_{k=0}^{6} \frac{k}{6} \cdot \frac{f(k,n)}{n^3}, \\ &= \frac{1}{6n^3}\big[6(n-2)^2 + 24(n-2) + 24\big], \\ &= \frac{1}{n}. \end{aligned}$$

Surprisingly, this is the same result as obtained earlier without detailed analysis of the groups of small cubes.

> All knowledge resolves itself into probability.
> —David Hume, *A Treatise of Human Nature*, 1985.

■ There are six red balls and eight green balls in a bag. Five balls are drawn randomly and placed in a red box; the remaining nine balls are placed in a green box. What is the probability that (the number of red balls in the green box) plus (the number of green balls in the red box) is not prime?

Answer. We denote by g the number of green balls in the red box. It is in the range $0 \le g \le 5$. Since the red box has five balls, $5-g$ of them must be red. This implies that in the green box (which has nine balls) there must be $8-g$ green balls and $g+1$ red balls (elementary but confusing). Thus, the sum (green balls in red box) + (red balls in green box) is $g + (g+1) = 2g+1$. This is an odd integer in the interval $[2, 11]$, and we notice that out of all the primes in this interval, only two, namely, 1 and 9, are odd.

They correspond to $g = 0$ and $g = 4$, respectively. The former is the case where all five drawn balls were red, and the latter is the case where four were green and only one was red.

The probability of drawing (zero green balls out of eight) and (five red balls out of six) our of a total of 14 balls is

$$p_1 = \frac{\binom{8}{0}\binom{6}{5}}{\binom{14}{5}} = \frac{6}{2002}.$$

Similarly, the probability of drawing (four green balls out of eight) and (one red ball out of six) out of a total of 14 balls is

$$p_2 = \frac{\binom{8}{4}\binom{6}{1}}{\binom{14}{5}} = \frac{420}{2002}.$$

Finally, the probability that either of these cases occurs is $1-(1-p_1)(1-p_2) = 0.212158$ and the probability that neither occurs is $1 - 0.212158 = 0.787842$.

■ Ten coins are labeled with numbers 1 through 10 on one side. They are tossed and the numbers on the top faces of the coins are added. What is the probability that the sum is at least 45?

Answer. The sum can be at most $1 + 2 + 3 + \cdots + 10 = 55$, so all we have to do is find out how many times ten distinct digits in the interval $[1, 10]$ can add up to 45 or less. However, it is easier to do the opposite, to figure out the (identical) probability that the sum is at most 10, i.e., the number of ways to add ten distinct digits in the interval $[1, 10]$ to get at most 10. This information is available as Sequence `A000009` of the OEIS, reference [Sloane 08], listing the number of ways to partition the integer n into distinct parts

n	0	1	2	3	4	5	6	7	8	9	10
	1	1	1	2	2	3	4	5	6	8	10.

Example: the integer 8 can be partitioned in 6 ways as follows

$$8 = 8 + 0 = 7 + 1 = 6 + 2 = 5 + 3 = 1 + 2 + 5 = 1 + 3 + 4.$$

The 11 numbers on the bottom row of the table add up to 43, so this is the number of ways to get at most 10 in our problem. Tossing the ten coins can produce one of 2^{10} results, which is why the probability in question is $43/2^{10} \approx 0.041992$.

> Who ever heard a theologian preface his creed, or a politician conclude his speech with an estimate of the probable error of his opinion.
>
> —Bertrand Russell.

■ Given a 3×3 grid, we place the numbers 1,2,...,9 in it at random.

1. What is the probability that the sum of every row and every column is odd?

2. What is the probability that the sum of every row, every column, and the two diagonals are also odd?

3. Similar, but exactly one of the diagonal sums is also odd?

Answer. Case 1. Each row, column, and diagonal consists of three integers, and the sum of three integers can be odd only if one of them is odd or all three are odd. There are only five odd integers in the interval $[1, 9]$, so in order for three rows to each sum to an odd value, one must have three odd integers and each of the other two must have a single odd integer, and similarly for the columns.

The row and column with the three odd integers have only odd integers, and must therefore have an odd integer at their intersection. There are $3 \times 3 = 9$ ways of choosing a row and a column. There are 5! ways (permutations) to choose the order of the five odd integers, and there are 4! ways of arranging the remaining four even integers. Accordingly, there are $9{\times}5!{\times}4!$ grids where each row and column sum to an odd number.

The total number of 3×3 grids in our case is 9!, which yields the probability

$$\frac{9{\times}5!{\times}4!}{9!} = \frac{1}{14} = 0.0714286.$$

Case 2. We already know that there are a row and a column, each having three odd integers. Therefore, there are not enough remaining integers to have three of them on a diagonal. The only way for the two diagonals to have odd sums is for each to have exactly one odd integer. Out of the nine ways of choosing a row and a column, only one, the one where the common integer is at the center (and therefore lies on both diagonals) guarantees that both diagonals will have an odd integer. Accordingly, the probability of this case is

$$\frac{5!{\times}4!}{9!} = \frac{1}{126}.$$

Case 3. In order to have one diagonal sum to an odd number and the other one sum to an even number, the odd integer that is common to a row and a column must lie on one of the four corner of our grid, which is why the probability of this case is

$$\frac{4{\times}5!{\times}4!}{9!} = \frac{4}{126}.$$

◇ **Exercise 7.15:** (A joke.) How can a stock broker obtain a large number of followers who eat out of his hand (at least temporarily) because of his unbelievably accurate predictions?

◇ **Exercise 7.16:** (1) After flipping a coin six times, what is the probability of getting exactly three heads (not necessarily in a row)?

(2) Tossing a coin six times, what is the probability of getting three heads in a row?

(3) what is the probability of getting exactly three heads in a row?

■ In the small town of Little Stempington there live two races, The Eloi (dark blue skin, constitute 90% of the population) and the Morlocks (dark gray skin, and only 10%). Neither race is especially violent, but there is some crime—mostly theft, burglary, and mugging—at which both races are equally guilty. Someone complains that he was

mugged by a Morlock in the dark. Tested by the police under similar lighting conditions, he manages to correctly identify both Eloi and Morlocks 80% of the time. The prevailing opinion in the town is that, given some suspects, the complainer has an 80% chance of identifying the culprit. It is your job, as a mathematician/statistician, to prove that his chance of doing so is only $4/13 \approx 0.31\%$.

The analysis is quite simple and is based on the very different sizes of the two races. The chance that the mugger is correctly identified AND is also a Morlock is $0.1 \times 0.8 = 0.08$. Similarly, the chance that the mugger is wrongly identified as Morlock but is an Eloi is $0.2 \times 0.9 = 0.18$. Morlocks are therefore identified $0.08 + 0.18 = 0.26$ of the time, but only 0.08 of the time are they identified correctly.

The chance of identifying a Morlock correctly is therefore 0.08/0.26 or 4/13.

■ Rounding a random fraction. Randomly, but with a uniform distribution, choose two real numbers x and y in the open interval $(0, 1)$. Divide them to obtain the real number $R = y/x$, and examine the integer I nearest R. We intuitively feel that I has the same probability of being even or odd, but experimentation (see the three-line Mathematica code below) as well as careful calculations [Weisstein-picking 07] show that the probability of I being even is 0.464602... instead of the expected 0.5.

```
n=10000.; ev=0;
Do[If[EvenQ[Round[RandomReal[]/RandomReal[]]],ev++],{i,n}]
ev/n
```

The video at reference [Morphocular 25] explains why this is so, and why the exact probability is $(5 - \pi)/4$. Thus, in addition to the surprising probability, we are also surprised (or are we) to see another occasion where π rears its head, whether beautiful or not. The explanation starts by pointing out that each of our random pairs (x, y) can be viewed as a point in the unit square, which is why the set of (x, y) pairs whose round quotient is 0, corresponds to a region in this square. From $0 < y/x < 1/2$ we conclude that $0 < y < x/2$, a condition which is satisfied by the points that constitute the blue region 0 of Figure 7.22a. The area of the square is unity and the area of this region is 1/4, which is why we can consider 1/4 the probability of obtaining a zero in our experiments.

Similarly, the probability that the round quotient of a random pair (x, y) is the even number 2 is the area of the blue triangular region 2. The condition for obtaining 2 is $3/2 \le y/x \le 5/2$ or $(3/2)x \le y \le (5/2)x$. The area of this region is one-half times its base $(2/3 - 2/5)$, times its height, which is 1. Thus,

$$\frac{1}{2}\left(\frac{2}{3} - \frac{2}{5}\right) = \frac{2}{15}.$$

Continuing in this way, the nth region is obtained by the condition

$$2n - \frac{1}{2} \le \frac{y}{x} \le 2n + \frac{1}{2}, \quad \text{or } (2n - 1/2)x \le y \le (2n + 1/2)x.$$

The areas of the regions are the individual probabilities that our rounded quotients will

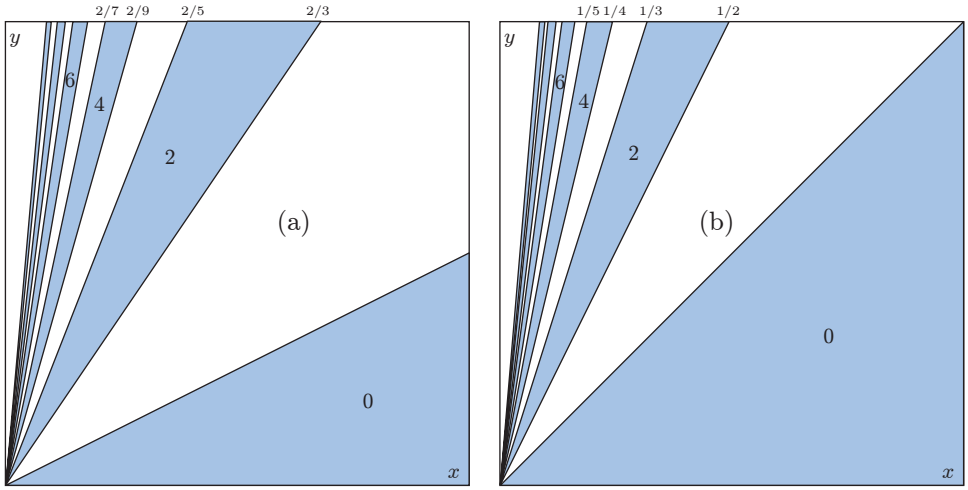


Figure 7.22: Regions in the Unit Square.

equal 0, 2, 4, 6, 8, and so on, ending with the infinite series

$$P = \frac{1}{4} + \frac{1}{2}\left(\frac{2}{3} - \frac{2}{5}\right) + \frac{1}{2}\left(\frac{2}{7} - \frac{2}{9}\right) + \frac{1}{2}\left(\frac{2}{11} - \frac{2}{13}\right) + \frac{1}{2}\left(\frac{2}{15} - \frac{2}{17}\right) + \frac{1}{2}\left(\frac{2}{19} - \frac{2}{21}\right) + \cdots.$$

Whose sum is the probability of obtaining an even integer in our experiments.

Once this sum is simplified to

$$P = \frac{1}{4} + \frac{1}{3} - \frac{1}{5} + \frac{1}{7} - \frac{1}{9} + \frac{1}{11} - \frac{1}{13} + \frac{1}{15} - \frac{1}{17} + \frac{1}{19} - \frac{1}{21}\cdots, \tag{7.8}$$

it would be recognized by experienced mathematicians as very similar to the Leibniz series of arc tangent, which is

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots.$$

In the special case where x equals 1 radian, this expression is reduced to

$$\arctan(1) = 1 - \frac{1^3}{3} + \frac{1^5}{5} - \frac{1^7}{7} + \cdots. \tag{7.9}$$

On the other hand, the tangent of $\pi/4$ (or 45°) is 1, which is why $\arctan(1) = \pi/4$.

After some derivations, Equations (7.8) and (7.9) produce the final result $P = (5 - \pi)/4 \approx 0.464602$. Another unexpected and unintuitive probability.

A variation. Instead of rounding our random fractions we can truncate them and compute the probability that the result is even. In the Mathematica code above we now change `Round` to `Floor`, and iterate 10,000 times to obtain 0.6497. An analysis similar to the one earlier, checking several even values, can easily explain this result. The condition

$\lfloor y/x \rfloor = 0$ implies $0 < y/x < 1$ or $0 < y < x$. Graphically, this corresponds to the blue region marked 0 in Figure 7.22b, and it adds 0.5 to the probability. The condition $\lfloor y/x \rfloor = 2$ implies $2 \le y/x < 3$ or $2x < y < 3x$. Graphically, this corresponds to the blue region marked 2 in the Figure, and it adds $0.5[(1/2) - (1/3)]$ to the probability. Similarly, the condition $\lfloor y/x \rfloor = 4$ implies $4 \le y/x < 5$ or $4x < y < 5x$, and it adds (the blue region marked 4) $0.5[(1/4) - (1/5)]$ to the probability.

In general, the condition $\lfloor y/x \rfloor = 2n$ implies $2n < y/x < 2n + 1$ or $2nx < y < (2n + 1)x$, and the total probability is therefore the infinite sum

$$P = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2} - \frac{1}{3}\right) + \frac{1}{2}\left(\frac{1}{4} - \frac{1}{5}\right) + \frac{1}{2}\left(\frac{2}{11} - \frac{2}{13}\right) + \frac{1}{2}\left(\frac{2}{15} - \frac{2}{17}\right) + \frac{1}{2}\left(\frac{1}{6} - \frac{1}{7}\right)\cdots$$
$$= \frac{1}{2}\left(1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \cdots\right).$$

Reference [Morphocular 25] shows in detail how to sum this series by means of a generating function, and proves that the mysterious 0.6497 obtained earlier is an approximation of $1 + \ln(2)/2 \approx 0.6534$. The surprise is that this time the probability depends on e, the base of the natural logarithms, while π does not appear.

> Every second that passes is a second lost forever, yet we always yearn for more. (Is this human nature or a paradox?)
>
> —H. G. Wells, *The Time Machine*, Heinemann, 1895.

7.11.1 Example: Real Roots of Quadratic Equations

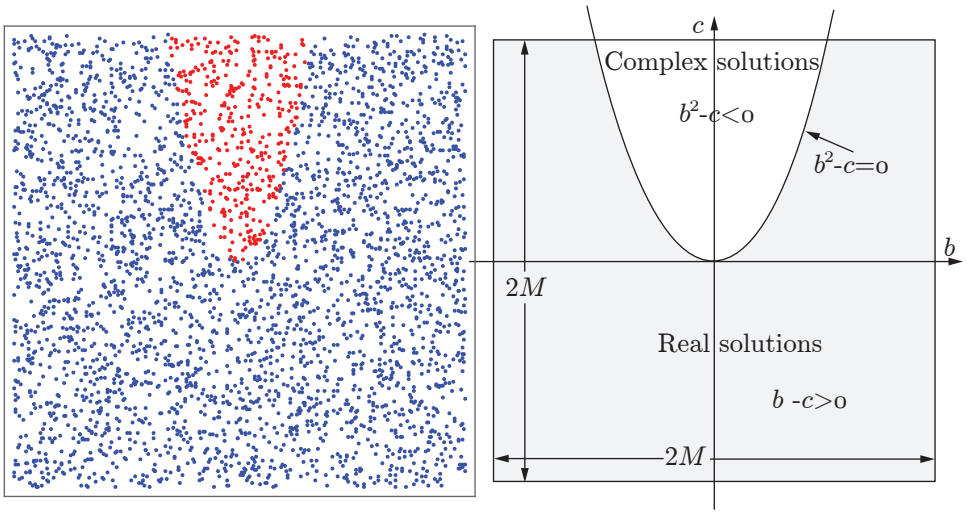
The general form of a quadratic equation is $ax^2 + bx + c = 0$, but since a cannot be zero, we divide by a and choose the special form $x^2 + 2bx + c = 0$. The task is to compute the probability that the solutions (roots) of this equation are real. As is well known, the solutions are real if the discriminant $\sqrt{b^2 - 4ac}$ is non-negative. In our case, the discriminant is $\sqrt{4b^2 - 4c} = 2\sqrt{b^2 - c}$, which reduces the problem to computing the area covered by the canonical parabola $c = b^2$. Since the parabola is an infinite curve, we start by restricting the problem to a square of side $2M$, and later let M grow without limit.

Figure 7.23 illustrates a canonical parabola $b^2 - c$ inside such a square with bc coordinates instead of the usual xy. The area inside the parabola corresponds to those coefficients b and c that produce complex solutions. It is obvious that this curve intersects the top of the square when $c = M$, which implies that $b = \pm\sqrt{M}$. The area of the parabola is therefore

$$\int_{-\sqrt{M}}^{\sqrt{M}} b^2 db = \frac{2}{3}M^{3/2}.$$

However, because we chose $2b$ instead of b as the coefficient of x, the area is actually $(4/3)M^{3/2}$, and it makes up a fraction of $1/(3\sqrt{M})$ of the total $4M^2$ square.

Basic Mathematica code is included with the figure, for readers to experiment with. It is easy to see that the probability values obtained in this simulation vary greatly depending on the values of n (number of simulation steps) and m (size of interval for random values).




```
e=0.;m=2;  n=3500; lc={}; rc={};
Do[{{b,c}=RandomReal[{-10^m,10^m},2],
If[c<=b^2,{AppendTo[lc,Point[{b,c}]],e++},AppendTo[rc,Point[{b,c}]]]},{i,n}]
Graphics[{Blue,lc,Red,rc},Frame->True, FrameTicks -> None]
e/n
```

Figure 7.23: Real (Blue) and Complex (Red) Solutions of a Quadratic Equation.

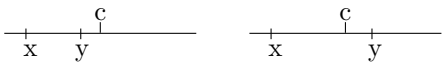
It is clear that as the square grows, the parabola (i.e., complex solutions) takes up less and less of the square's area, which implies that the probability of our equation having real solutions grows steadily with the interval of the random coefficients b and c. If these coefficients can be any real numbers, the probability is 1; no complex solutions.

Note. In order to simplify the derivation, we have divided the original coefficients a, b, and c of the equation by a. The trouble is that if the original b and c are independent and are uniformly distributed over a certain interval, the new coefficients b/a and c/a are not independent and neither are they distributed uniformly.

7.11.2 Probability that Three Segments Form a Triangle

Take a straight segment of length one, and make two random cuts to produce three shorter segments. What is the probability that those segments can form a triangle?

Solution 1. A 1-unit segment cut into three shorter segments can form a triangle if all three are smaller than 1/2 unit. The diagram shows such a segment with c as its midpoint, and two cuts at x and y. If both cuts are on the same side of c (the probability of which case is 1/2), then one segment must be longer than 1/2 unit, and there is no triangle.



The probability of the two cuts being on different sides of c is therefore also 1/2, but again the diagram shows that in half of those cases (whose probability is 1/4) the

distance between the cuts is longer than 1/2 unit. Thus, a triangle is possible only if the two cuts are on different sides of c and the distance between them is less than 1/2 unit. The probability of this is 1/4.

Solution 2. Viviani's theorem claims that for any point P inside an equilateral triangle, the sum of the distances from P to the three sides of the triangle is constant and equal to the triangle's altitude.

Start with the equilateral triangle of altitude 1, Figure 7.24a, and choose a point at random inside it. The three red segments are the distances of the point from the triangle's edges, and their sums is always one unit (the altitude of the triangle). The locus of the points where all the distances are less than 1/2 is the smaller triangle, in dash, and its area is 1/4 that of the larger triangle. Simple!

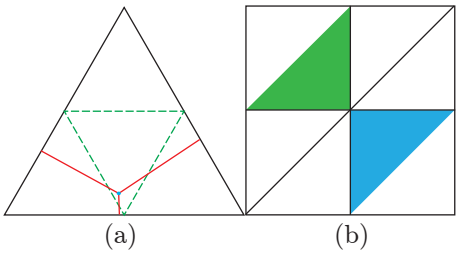
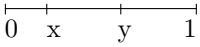


Figure 7.24: Segments to Form a Triangle.

This is a simple solution, but there still should be proved rigorously that moving the point inside the triangle produces the same distribution of three lengths as the one obtained when cutting a segment into three parts.

Solution 3. The first diagram shows a possible partition of the one-unit segment by making cuts at points x and y such that $x < y$. The lengths of the segments are x, $y - x$, and $1 - y$.

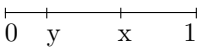


The condition for forming a triangle is for each of the three segments to be shorter than the sum of the lengths of the other two. This can be expressed as

$$\begin{aligned} x + (y - x) > 1 - y &\to y > 1/2, \\ x + (1 - y) > y - x &\to y < x + 1/2, \\ (y - x) + (1 - y) > x &\to x < 1/2. \end{aligned} \tag{7.10}$$

Figure 7.24b illustrates the graphical meaning of these conditions. The first condition $y > 1/2$ means the top half is the unit square. Condition $x < 1/2$ means the left half of the square, and condition $y < x + 1/2$ means the area above the 45° (dashed) line in Figure 7.24b. These define the triangular green area.

The next diagram implies that we still need to consider the case where the two cuts x and y change places.



◇ **Exercise 7.17:** Based on the above derivation, analyze this case.

◇ **Exercise 7.18:** A rod is randomly broken into two parts. What is the average length of the smaller part?

7.11.3 Throwing Dice

How many times, on average, do you have to toss a die in order to get a 6?

Approach 1, intuitive. We know from experience that the probability p of getting a 6 (or any other integer between 1 and 6) is 1/6. Therefore, we can expect one toss out six to produce a 6, implying that the average number m of die tosses to get any result is six.

Approach 2, mathematical. If p is the probability of getting 6 in any throw, then $q = 1 - p$ is the probability of getting a different result. This leads to the following conclusions:

- In one throw, the probability of success is p.
- In two throws, the probability of success is pq, because the first throw failed and the second one succeeded.
- In three throws, similarly, the probability of success is pq^2.
- In n throws, we failed $n - 1$ and succeeded in the nth time, so the probability is pq^n.

These probabilities should add up to unity, so we check

$$p + pq + pq^2 + \cdots = p(1 + q + q^2 + \cdots) = \frac{p}{1-q} = \frac{p}{p} = 1.$$

[It is valid to use the sum of a geometric series because q, as a probability, must be in the interval $[0, 1]$, and the special case $q = 1$ implies $p = 0$, so is trivial.]

To find out how many die throws are expected before getting a certain outcome, we use the statistical expectation E. Here is what this term means. Given a process X (formally, a random variable) whose outcomes (or values) can be x_1, x_2, x_3, ..., x_k, with probabilities p_1, p_2, p_3, ..., p_k, the expectation $E(X)$ is defined as the weighted sum of the outcomes x_i, weighted by their probabilities p_i. Thus

$$E(X) = x_1p_1 + x_2p_2 + x_3p_3 + \cdots + x_kp_k.$$

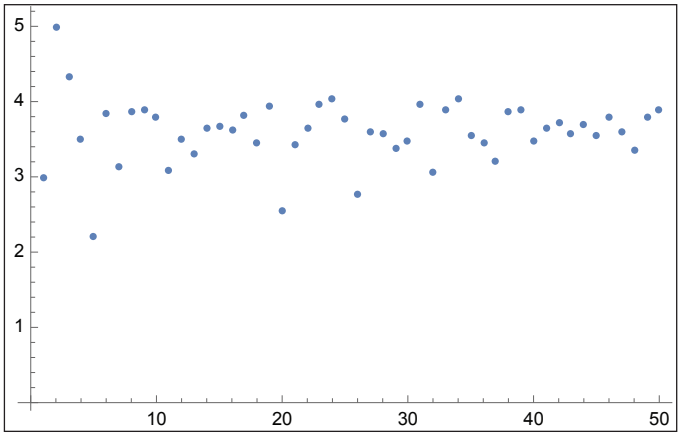
This definition makes sense since the sum of the probabilities is unity, implying that expectation is a barycentric sum (Page 470).

In the special case of uniform probabilities, where $p_i = p = 1/k$, the weighted sum of the expectation is reduced to the arithmetic average (Section 10.8) $(\sum_i x_i)/k$, which is why we can claim that expectation is more general than average because it takes into account the fact that some outcomes are more likely than others.

Getting back to die throws, the expectation is the weighted sum

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5,$$

As illustrated by the simulation of Figure 7.25. Because this is a case of equiprobable probabilities, the 3.5 expectation is the expected average of die throws, and a simulation, real or digital, always confirms this result, which is also supported by the law of large numbers.




```
lc={};
Do[{e=0;Do[e=e+RandomInteger[{1,6}],{i,1,k}],AppendTo[lc,e/k]},{k,1,50}]
ListPlot[lc]
```

Figure 7.25: Simulating Dice Throws.

Throwing a die several times is a process (or a random variable) with outcomes whose probabilities depend on the number of throws. The expectation of this process is therefore the sum

$$E = p + 2pq + 3pq^2 + 4pq^3 + \cdots \tag{7.11}.$$

Multiplying the two sides by q yields

$$qE = pq + 2pq^2 + 3pq^3 + 4pq^4 + \cdots, \tag{7.12},$$

and subtracting Equation (7.12) from Equation (7.11) produces

$$E - qE = p + pq + pq^2 + pq^3 + \cdots,$$

which reduces to $E(1 - q) = 1$ or

$$E = 1/p. \tag{7.13}$$

In our search for the average number of throws to obtain a 6, the probability of success is $p = 1/6$, which results in $E = 6$, as expected.

7.12 Probabilistic Counting and HLL

HLL stands for Hyper-Log-Log. It is the latest improvement of an algorithm which originated in 1985 with reference [Flajolet 85]. The HLL algorithm computes a fast estimate of the number of distinct items in a dataset. The term probabilistic counting is used with HLL, because this algorithm makes use of basic probability concepts.

We start with simple illustrations of the problem. Imagine a train or subway station in a big city. Thousands pass through the station every day, and many may find themselves there several times a day. The transportation authorities are interested in how many distinct individuals use the station daily. A similar example may be a free, historically-important exhibition (Picasso, Vermeer, and Rembrandt come to mind) in a large museum or exhibition hall. The organizers want to know, at least approximately, how many distinct individuals have visited during the exhibition. The exhibition being free, the organizers may also be interested in how many times the average visitor came, but this is a separate problem.

It is possible to place face-recognition cameras at the entrances, to identify individual faces and count distinct visitors. Currently, in 2024, this technology is not mature and may provide only a rough estimate of the count, but the HLL technique outlined here promises better estimates at the cost of simple computations.

◇ **Exercise 7.19:** Come up with another example of a useful application of counting distinct events (or visitors).

The term "cardinality" of a set of items is used to indicate the number of distinct elements of the set.

Given a large set of items (in practice, a file on a disk drive or other digital volume) the simplest way to estimate its cardinality is based on sampling. Given a file of size N, take a small random sample of size N_0, compute its cardinality using any algorithm that counts and compares items directly, and consider the ratio N/N_0 an estimate of the cardinality of the entire file. This is simple, but less accurate than the HLL algorithm. Also, a random sample may not be a good sample if different regions of the file exhibit different distributions of its data.

The HLL algorithm is based on probability concepts, so we start with flipping coins. The probability of a run of n consecutive heads (or tails) while flipping a coin n times is $1/(2^n)$. For $n = 5$, this is $1/32$, a small, but not very small, number. We can expect to obtain such a run by actually flipping coins for perhaps an hour or so. However, the probability of getting 20 heads in a row is the much smaller $1/2^{20} = 1/1{,}048{,}576 \approx 0.000000954$, so it may take several years of actual flipping to finally see such a run.

Turning this conclusion on its head, imagine that we flip a coin and suddenly see a run of 20 heads. This very probably means that we must have been flipping coins for a long time. In general, this idea implies that while flipping coins, the length of the longest run of heads that we have seen so far is an excellent indication of the time we have been spending (wasting?) in flipping coins.

This simple, surprising insight into probabilities and coin flipping is at the heart of the HLL algorithm, but since any practical implementation of the algorithm will be on a computer, the algorithm must be described by means of numbers (specifically, binary numbers) instead of with coins. As the first step in this description, we go back to

our museum and its many visitors. In order to introduce a numeric element into the discussion, we assume that visitors are identified not by their faces, but by the six (or seven or eight) least-significant digits of their telephone number. We must assume that those digits are uniformly distributed and that each of the six digits of a number may, with equal probability, be any of the ten decimal digits.

Once we obtain those six digits from each incoming visitor, we simply count the number N of leading (most-significant) zeros of each number and only save the largest N seen so far. (The least-significant digits can be used instead.) When a visitor arrives at the museum for a second time, his number is again checked, but since it had already been checked and was used to update (or not update) N, the second appearance of this number is essentially ignored. The earlier discussion of coin flipping implies that, in random data, a run of k zeros will occur on average once in every 10^k items. Thus, if N stands at, say, 4—implying that one or more visitors with numbers of the form $0000xx$ have been found—chances are that about $10^4 = 10{,}000$ visitors have already entered and their numbers checked.

Is it possible that all 10,000 visitors had numbers of the form $0000xx$? Not if the numbers are distributed uniformly. If we suspect that visitors' telephone numbers are somehow related, perhaps because of their being from the same city, then another measure must be chosen as a key to the HLL algorithm. However, reference [Flajolet 85] and all its successors propose that a hash table should be used to (1) improve the uniformity of whatever keys are used and (2) guarantee that only distinct keys are counted. (The latter is true because the hash table generates identical keys for identical inputs.) Such a table generates binary numbers, which is why any logarithms used by HLL must be to base 2. We now pause the description of HLL in order to introduce the important concept of a hash table and hashing in general.

Hashing and hash tables

The JAVA programming language is commonly used in computers and other programmable devices. Among others, JAVA is the main tool used to implement the well-known Android operating system. A JAVA compiler is software that reads a JAVA program and converts it to a set of machine instructions. One of first tasks of this and other compilers is to identify the keywords of JAVA in the input program and invoke the correct routine to deal with each. Typical keywords in a higher-level programming language are phrases such as `public`, `do`, `int`, `if`, `else`, and `while`. When the compiler identifies such a phrase, it uses a hash table to convert the phrase to an index, a location within an array A. Each location of A contains the start address of a special routine that reads and compiles the particular JAVA text that is associated with the phrase.

The important feature of this type of hashing is collision avoidance. Different phrases should hash to different indexes. Reference [WikiHash 24] lists several methods for handling hash collisions.

Another example of hashing is password verification. We all have many passwords that permit us to use various online accounts and resources. When a new user opens an account at, say, Bobmart, the Bobmart computer asks for a username and a password, a string of characters. It then scrambles the bits that constitute the password (a process known as hashing) and stores the resulting binary string, together with the username, in a password file. When the user logs into Bobmart's computer, they enter their name

and password. The Bobmart computer (1) hashes the password to create a string C, (2) uses the username as an index to the password file, and (3) compares C to what it finds in that location.

In this type of hashing, it is crucial to have an irreversible (one way) computational process to create string C. If a hacker manages to break into Bobmart's computer and steal the password file, they may be unable to use any of its hashed C strings in order to retrieve an original password.

Thus, a hash table is a set of rules or computations that starts with a string of characters and generates a string that may then be used as an array index or simply stored for future verification. The hashing computations must be irreversible (one way) and should avoid collisions as much as possible, i.e., they should hash different inputs to different outputs. Also, every bit in a string C that is output by hashing should have equal probability of being a zero or a 1.

(End of hashing and hash tables.)

The HLL algorithm starts by hashing an input data v into a binary string x, but before we continue with the algorithm it is important to analyze the rare case where one of the early x strings output by the hash process has many consecutive leading (or trailing) zeros. The statistical term for such an occurrence is outlier, but in English, such an unfortunate case is often termed a fiasco.

> Fiasco, a thing that is a complete failure, especially in a ludicrous or humiliating way.
> —From the dictionary.

The developers of HLL have adopted the following elegant solution to this problem. Create an array M of m counters (or registers) that are initialized to 0. Any string x generated by the hash process is partitioned into two parts, an index j to array M and a remainder w whose run length $r(w)$ of trailing zeros is measured. If $r(w)$ is greater than the value of $M[j]$, then $M[j]$ is set to the new maximum $r(w)$. An outlier w may still happen and may greatly impair the maximum stored in a counter, but outliers are rare and if the number m of counters is large enough, their average may considerably mitigate the damage done by any outlier.

Thus, at the end of the input stream, when all the data items have been read and processed, the algorithm computes the average of the m counters of array M and multiplies the product by m to obtain the final result, the estimate of distinct data items. The only remaining question is what type of average to use in order to minimize the impact of the few very large counters. The best choice of average, as explained in Section 10.8, is the harmonic mean.

It is easy to see that a large number m of counters increases the accuracy of the final estimate, while only slightly increasing the memory requirements of array M. However, a large m requires a large index j, which leaves fewer bits in string x for the all-important remainder w with its consecutive zeros.

In practical applications, the input file may be very large, consisting of billions of data items. A simple way to significantly increase the speed of HLL is to run this algorithm on the p individual processors of a large, parallel computer. Each processor reads part of the input file, runs the algorithm, and ends up with its individual m counters. When all the processors are done, the counters are merged by a union operation. Counters #1 of the p processors are examined and the largest of their values is selected. This

is repeated for the remaining $m-1$ counters. HLL is therefore an example of a parallel algorithm.

Following is a summary of the algorithm. A data item v is input. It is hashed to a bitstring x, which is partitioned to form an index j and a remaining part w. In order for j to be an index to array M of m counters, its length should be $\log_2 m$. Counter $M[j]$ is compared to the number $r(w)$ of consecutive zeros (leading or trailing) in w. If $r(w) > M[j]$, $M(j)$ is set to $r(w)$.

When the entire input file has been read and processed, the harmonic average of M is computed by

$$Z = m \sum_{j=1}^{m} \frac{1}{2^{M[j]}},$$

and the final estimate of the cardinality of the input is mZ. In practice, this estimate suffers from the inevitable hash collisions, so the developers of HLL recommend to further multiply mZ by the complex and hard-to-compute constant α_m which is given by

$$\alpha_m = \left[m \int_0^{\infty} \left(\log_2 \left[\frac{2+u}{1+u} \right] \right)^m d\,u \right]^{-1}.$$

In initial experiments with very large sets of data, HLL produced estimates that were about 2% off the correct values.

It only remains to explain the game of the name. Why Hyper-Log-Log?

This algorithm went through several development steps. The first version was named LogLog because after processing all n data items, the total number of bits in the m counters is of the order of $\log_2(\log_2 n)$. Here is a detailed explanation of this fact.

First, we discuss the length $L(k)$ (the number of digits) of a given integer k. The length depends on the size of k, and the following discussion shows why it is expressed by a logarithm.

Each of the first 10 decimal digits 0–9 is one digit long. Each of the first 100 decimal integers 0–99 is two digits long. It doesn't take lone to realize that each of the first 10^d decimal integers is d digits long. Thus, the length of the integer k is $\lceil \log k \rceil$ decimal digits. Since there is nothing special about 10 as the base of a numbering system, the same argument applies to binary numbers, and the length of k in bits is also a logarithm, but to base 2.

Now, to the choice of the name LogLog. What is actually stored in a counter $M[j]$ of M is the length $r(w)$ of a run of consecutive zeros in the output string x generated by the hash function for a visitor. The basic idea that underlies HLL is that the length of the longest run of heads that we have seen so far is an excellent indication of the number of coin tosses so far. Therefore, if we expect a maximum of n visitors, the largest value of $r(w)$ can be about $\log_2 n$.

On the other hand, since HLL deals with estimates, we don't need to know the precise value of $r(w)$. Instead of storing $r(w)$ in $M[j]$ we can store its length in bits. For example, if $n = 100{,}000$, then we can expect the maximum $r(w)$ to be $5 = 101_2$, which is expressed in three bits, so the number 3 is what eventually can be saved in counter $M[j]$.

This discussion shows that the total number of bits needed to save the various maxima in the m counters of array M is essentially $\log_2 \log_2 n$, hence the name LogLog.

In the second version of HLL, dubbed SuperLogLog, the problem of outliers was dealt with by deleting several counters with the largest counts, while retaining the counters with 70% smallest values.

The third version, assigned the name HyperLogLog, uses harmonic instead of geometric average.

Poem on Probability and Chance

In life's vast weave of fate and chance,
A dance of numbers, randomness,
Probability's nuanced glance,
A subtle play in life's vastness.

In coins that spin and dice that roll,
In cards dealt from an unseen hand,
The likelihood of outcomes' toll,
Unfolds within a structured land.

A flip of heads, a toss of tails,
In each event, a story starts,
The odds and ends of life's details,
In randomness, a beating heart.

The bell curve's arc, a graceful rise,
In standard forms, predictions lie,
Normality where data lies,
A realm where chaos seems to die.

In quantum worlds, uncertainty,
Where particles in mystery,
Exist in states we cannot see,
Yet bound by chance's history.

From weather's whims to stock market trends,
From simple games to cosmic ends,
Probability shapes and bends,
The path of fate, the dice it sends.

Each lottery ticket, a distant dream,
In odds so vast, the chances gleam,
A spark of hope, a fleeting beam,
In numbers' game, life's silent scheme.

The gambler's fall, a siren's call,
In past events, no future claim,
Yet patterns seem to rise and fall,
In streaks and runs, the mind's own game.

In science and in daily life,
Probability cuts like a knife,
Through certainty and through the strife,
It marks the boundaries of our life.

Bayesian thoughts and Pascal's bet,
In reason's light, our paths are set,
With every choice, a thread is met,
In life's great web, a net is kept.

The dice are cast, the wheels will spin,
In chance's realm, we seek to win,
Yet often find, within, akin,
A dance of fate and luck within.

So in the chaos, find the rhyme,
In odds and ends, in space and time,
Probability's grand paradigm,
A dance of chance, a silent chime.

Each moment is a coin's soft spin,
Each breath a roll, where odds begin,
In probability, we find our kin,
In life's grand game, we play, we win.

By `chatGPT 4o`, May 2024.

The 50-50-90 rule: Anytime you have a 50-50 chance of getting something right, there's a 90% probability you'll get it wrong.
—Andy Rooney.

8
Geometry

Geometry Quotes

Let no one ignorant of geometry enter.

—Engraved at the door of Plato's Academy

There is geometry in the humming of the strings, there is music in the spacing of the spheres.

—Pythagoras

I was very slow in maths, geometry I actually enjoyed.

—Liam Neeson

Music is the arithmetic of sounds as optics is the geometry of light.

—Claude Debussy

Geometry is not true, it is advantageous.

—Henri Poincare

There is no royal road to geometry.

—Euclid

The description of right lines and circles, upon which geometry is founded, belongs to mechanics. Geometry does not teach us to draw these lines, but requires them to be drawn.

—Isaac Newton

One of the great sources of employment for people with Ph.D.s in geometry is the animation industry.

—Margaret Wertheim

It is hard to convince a high-school student that he will encounter a lot of problems more difficult than those of algebra and geometry.

—E. W. Howe

You can't criticize geometry. It is never wrong.

—Paul Rand

Poetry is as precise a thing as geometry.

—Gustave Flaubert

The early study of Euclid made me a hater of geometry.

—James Joseph Sylvester

Geometry was the first exciting course I remember.

—Steven Chu

I've always been passionate about geometry and the study of three-dimensional forms.

—Erno Rubik

It is marvelous enough that man is capable at all to reach such a degree of certainty and purity in pure thinking as the Greeks showed us for the first time to be possible in geometry.

—Albert Einstein, as quoted by Paul Arthur Schilpp, 1951.

8.1 Fractals

The term "fractal," from the Latin fractus, meaning broken or fractured, was coined by Benoît Mandelbrot in 1975. Mandelbrot used this term to describe geometric shapes that are extremely complex, and in his writings and lectures he managed to popularize it, perhaps because fractals often have intriguing and surprising shapes, thereby serving as a form of art, especially when artificial colors are added.

We start with an informal definition that is easy to understand. A fractal is a geometric shape that is either self similar, or parts of it are self similar, or it scales differently from most other geometric shapes, or is infinitely jagged and rough (non differentiable). There have been attempts to come up with a rigorous definition of these intriguing objects. A widely accepted definition is: A fractal is any object whose fractal dimension exceeds its topological dimension. Fractals are also found in processes that develop over time, and this fact must also be included in any complete, rigorous definition. For the purposes of this section, however, it is enough to consider the three parts of the definition above.

- Self-similarity. When a circle is magnified, we may end up seeing only a small part of it, and this part starts resembling a straight line as the circle grows steadily. Thus, a circle, and many other familiar geometric figures, are not self-similar. However, a

straight segment, a square, and a cube are self-similar; they look identical to smaller parts of themselves. Hence, self similarity is not a sufficient condition for a fractal.

- Scaling. When a circle's radius is doubled, its circumference also doubles, but its area increases by four. Similarly, when the radius of sphere is doubled, its volume increases by eight. Fractals, however, behave differently in this respect. Imagine an n-dimensional complex geometric shape that consists of infinitely many straight segments and occupies an area or a volume in that space. If the length of every segment is doubled, the space occupied by this object increases by a factor that is not necessarily an integer. This factor is the fractal dimension of the object, and is normally different from its topological dimension, which is n, the number of dimensions of the space where the object "lives."

- The familiar functions are either smooth (have a well-defined slope or tangent vector at every point) or have a number of points, finite or infinite, where they are non-differentiable because of a corner or a sudden turn. A fractal, on the other hand, may be continuous, but may have a corner or a change of direction at every point in its codomain. It may be infinitely jagged.



A Fractal in the shape of a frog, by chatGPT 5.

The fractal dimension is the standard mathematical tool used to define fractals. Traditionally, dimensions are natural numbers, but fractal dimensions tend to be fractions. When a straight segment of length r is broken into two identical segments, each remains one-dimensional and has length $r/2$. Similarly, partitioning a square up into four smaller squares preserves the two-dimensionality and cuts each side to $1/2$ of the original. Breaking a cube produces eight smaller cubes that are still three-dimensional, but with side length of $1/2$ of the original. Thus, length sizes change in a predictable way regardless of the dimension of the figure.

In contrast, the area or volume of an object (we'll use the term mass instead of area or volume, since geometric objects may have many dimensions) varies according to the spatial dimension of the object. Assuming that the three geometric figures above are real and have mass, what happens when they are reduced in size as before? The mass of a

half straight segment is reduced by a factor of 1/2, which can be written as $(1/2)^1$. The mass of each of the four smaller squares is $1/4 = (1/2)^2$, and that of each of the eight smaller cubes is $1/8 = (1/2)^3$. The mass (and with it, the area or volume) decreases according to the spatial dimension of the figure, and those dimensions are integers.

It is therefore easy to agree that this also applies to higher dimensions. If a symmetric n-dimensional figure would be broken up into 2^n smaller copies of itself, the mass of each would be smaller than the original mass by a factor of $(1/2)^n$.

We can summarize these conclusions in (1) a length L varies according to $(1/2)L$, but (2) a mass M varies by $(1/2)^d M$, where d is the spatial dimension of the figure.

Now take a familiar self-similar fractal, such as the Sierpinski triangle, Figure 8.4. It consists of three main parts, so it makes sense to partition it into three parts. From the figure, it is clear that the length of each part is 1/2 the original length. However, because it is made of three parts, it is reasonable to claim that the "mass" of each would be 1/3 the "mass" of the original triangle. The earlier discussion of mass suggests that its mass would shrink by a factor of $(1/2)^d$, which yields the important equation

$$\left(\frac{1}{2}\right)^d = \frac{1}{3}, \Rightarrow 2^d = 3, \Rightarrow d = \log_2 3 \approx 1.585.$$

This non-integer, fractal dimension makes sense because the Sierpinski triangle is defined as the limit of infinite recursion where, in each step, the figure consists of a curve passing through many points, but leaving many other points vacant. The curve never becomes a solid surface, and its dimension is always less than 2.

It seems that neither the length of this curve nor its area are good indications of its dimension. In the limit, its length is infinite, and its area is zero, because the figure is always a curve consisting of straight segments connecting points, never a solid surface. On the other hand, the figure of 1.585 suggests that the "final" figure (in practice, of course, it is never final) is more than a mere curve, yet not exactly a two-dimensional area. Some may suggest that perhaps at the limit, this triangle is 58.5% of the way from being a curve to being an area.

Another example of a fractal dimension is the von Koch snowflake, Figures 8.3 and 8.11. This fractal is naturally broken into four identical copies of itself, each scaled by a factor of 1/3. Hence, the length (or in general, the scaling factor) is 1/3 and the mass scaling factor is 1/4, leading to the relation $(1/3)^d = 1/4$, that solves to $d = \log_3 4 \approx 1.262$. This lower fractal dimension makes sense. Comparing the shapes of the von Koch and Sierpinski curves, it seems intuitively that even at the limit, the former occupies less space on the xy-plane than the latter.

The next example is the Hilbert curve, Figure 3.63, which can also be viewed as a version of the von Koch curve, right angled. Here, the scaling factor is 1/4 and the mass scaling factor is 1/8, leading to the relation $(1/4)^d = 1/8$, which solves to $d = \log_4 8 = 1.5$, an exact value which suggests that this fractal is somehow midway between a curve and a two-dimensional surface.

Geometric shapes, which are precise and well-defined, have fractal dimensions that are easy to calculate, but what about an arbitrary shape whose boundary may be rough and jagged, as in Figure 8.1. The figure resembles an island, whose coastline is continuous, but is fractured and is not described by a mathematical function. Even more. As

we zoom in and examine smaller details of this coastline, it often happens that what we see is always about the same. The coastline expands before us. We see less of it, but in more detail, yet this detail is similar overall to the way the entire coastline looked earlier. If we observe such behavior, then the coastline is self-similar and is a fractal.

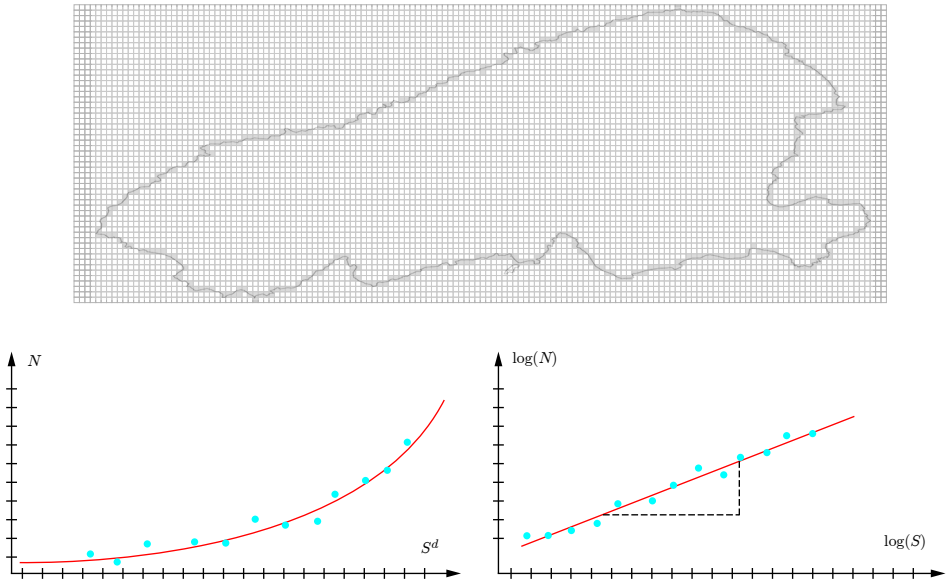


Figure 8.1: A Fractal Coastline.

In order to determine if such a coastline is self-similar, we surround the entire island by a screen and mark those screen boxes that cover the coastline. The number N of boxes provides a rough approximation of the unknown length of the coastline. In order to get a better approximation, we shrink the screen by a factor S, and again mark the boxes that cover the coastline. Each screen box is now smaller, so more boxes cover the coastline. Their number is therefore a better approximation of the length of the coastline.

It is reasonable to assume that if the coastline is self-similar, the number N of boxes would grow proportionally to the shrinking scale factor S. It is even more general to expect N to be proportional to a certain power d of S, where d depends on the particular fractal nature of the coastline.

We therefore write $N \propto S^d$ or $N = cS^d$, where c is a proportionality factor. After scaling the screen many times and counting many values of N, we can try to estimate d by plotting one point with coordinates (S, N) for each scaling, and trying to fit a parabola of the form $f(x) = cx^d$ to the points as depicted in the figure. A better approach is to plot the points on a log-log scale $(\log S, \log N)$, which in principle, if the coastline is self-similar, should give a straight line of slope d. In practice, the points would normally be located close to this line, and a method such as least squares (Section 8.5) can then be applied to compute the line closest to all the points, to measure its slope, and find

d, the fractal dimension of the coastline.

Historical Notes

The scientific recognition of fractals seems to have started with Gottfried Leibniz who mentions the term "fractional exponents" in his writings. In 1872, Karl Weierstrass published the first description and definition of a continuous function that is non-differentiable everywhere (Section 8.2). As is common when a revolutionary effect or concept is first announced, some mathematicians couldn't accept such functions. They started referring to them as monsters and declared that the entire discipline of mathematics was in a crisis.

However, later in the 19th century, Georg Cantor introduced what is now called the Cantor sets, and was followed by Felix Klein and Henri Poincaré who both introduced a family of fractals that came to be called self-inverse fractals. Those pioneers were followed by (among others):

(The Cantor set, Figure 8.2, is obtained from [0,1] by removing the open middle third and iteratively removing the middle third of each remaining interval.)

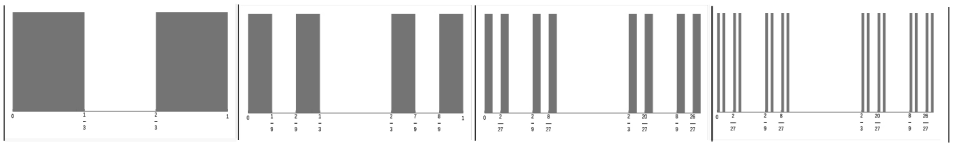


Figure 8.2: The Cantor Set.

- The 1915 Koch snowflake of Niels Helge von Koch.

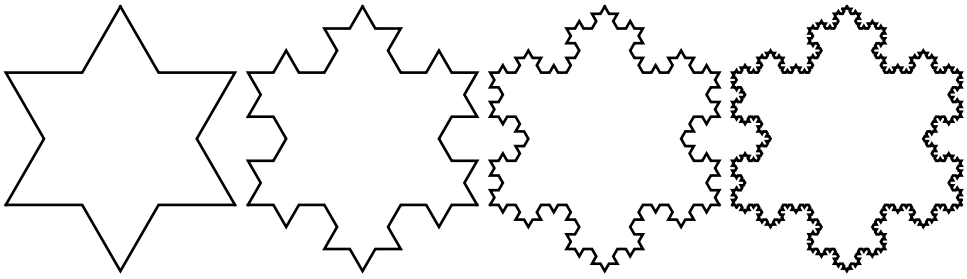


Figure 8.3: A Koch Snowflake.

- The 1915 Sierpinski triangle, Figure 8.4, and Sierpinski carpet (or gasket), both by Wacław Sierpiński.
- The 1918 Julia set, by Gaston Julia and, independently, Pierre Fatou.
- The 1938 Lévy C curve, named after Paul Lévy (Figure 8.8). This name, by the way, is a classic example of an object named after the last person who invented it,

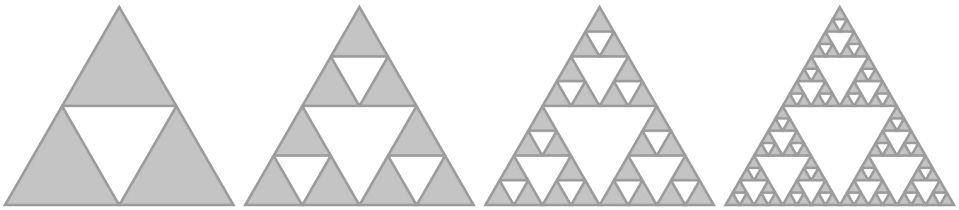


Figure 8.4: The Sierpinski Triangle.

because both Ernesto Cesàro in 1906 and Georg Faber in 1910 described and analyzed this object before Lévy.

- The name of M. C. Escher is often mentioned in connection with original and unusual drawings and paintings, and it seems that Escher was also a pioneer in fractal art. His well-known 1959 drawing *Circle Limit III* consists of shapes that shrink as they are repeated and approach the edges of the drawing, thereby creating a circular self-similar pattern that looks the same when magnified.

- We eventually reach Benoît Mandelbrot, the individual primarily credited for popularizing fractals among mathematicians, scientists, artists, and the public. He became interested in the subject during the 1960s, focusing on self-similarity in his thoughts and writings, but his most significant impact on the subject's popularity came from his famous book *The Fractal Geometry of Nature*, published by W. H. Freeman in 1977. Of the numerous individuals who were captivated and intrigued by fractals, I will only highlight Loren Carpenter, whose software for creating and displaying fractal landscapes was unexpected and a revelation to many.

Techniques for generating fractals

- Iterated function systems (IFS), subsection 8.1.1. This type of fractal starts with a simple geometric figure (typically two- or three-dimensional) such as a set of points, segments, or arcs—that is later refined in steps where it is scaled, and smaller copies of itself are combined after being moved, rotated, or reflected. This process normally results in a self-similar figure. Figure 8.5 illustrates a simple example.

  Reference [Hutchinson 81] is the original proposal for this type of fractals, but it was [Barnsley 13] that help popularize it.

> IFSs provide models for certain plants, leaves, and ferns, by virtue of the self-similarity which often occurs in branching structures in nature.
>
> —Michael Barnsley.

Some familiar IFS fractals are the Koch snowflake, Cantor sets, Haferman carpet, Peano curve, Harter-Heighway dragon curve, T-square, Menger sponge, and the Sierpinski triangle, carpet, and gasket.

The definition of an IFS is simple. It consists of the initial shape (points, lines, and other basic geometric figures), the scaling factor, and rules for combining the copies in the construction steps. These items require only a small storage space, which begs the question of the inverse IFS problem. Given a digital image, a set of pixels, find an IFS

Mandelbrot obituaries

[I]f we talk about impact inside mathematics, and applications in the sciences, he is one of the most important figures of the last fifty years.

—Heinz-Otto Peitgen.

An icon who changed how we see the world

—Chris Anderson.

Mandelbrot had a powerful, original mind that never shied away from innovating and shattering preconceived notions [. . . h]is work, developed entirely outside mainstream research, led to modern information theory.

—Nicolas Sarközy.



Here it is, David, an original Mandelbrot portrait where his suit becomes a canvas for fractal terrains. Mountains rise from lapels, valleys swirl through the folds, and spirals echo the infinite recursion of his own image nested within itself. It's a fusion of identity and geometry, as if the very fabric of his being is stitched from the landscapes he revealed to the world. A fractal homage to the man who saw infinity in geometry.

Image and description by MS Copilot, Sept. 2025.

(initial shape and rules) that generates this set, exactly or approximately, after several iterations. Such an IFS is a compressed version of the original image, a fact which has prompted researchers to develop methods for finding an IFS for a given image.

■ Strange attractors. Mathematics is a vast field of knowledge, encompassing a large number of areas or disciplines. One such area is the study of dynamical systems. Such a system is a set of mathematical points or physical objects that move in space and time. Examples of dynamical systems are the movement of a mathematical point along a curve, the apparently random motion of air or water molecules, the periodic motion of a pendulum, and the variations of prices, interest rates, unemployment rates, and stocks/shares in the economic world.

A dynamical system evolves, develops, and changes its state, shape, and form over time. It may happen that at a certain point in time, the states of a system start approaching a limit, where the system remains stationary. Such a limiting state is referred to as an attractor or a stable state. When a system at this state is slightly disturbed, it tends to return by itself to the attractor state.

An attractor can be a point (in space or in time), a finite set of points, or a surface (an infinite set of mathematical points embedded in three-dimensional space). An attractor can also be a complex, infinite set of points that can be considered a fractal because it may be assigned a non-integer dimension. Such an attractor is called strange.

In practice, a strange attractor can be the limit of a nonlinear process. Two well-

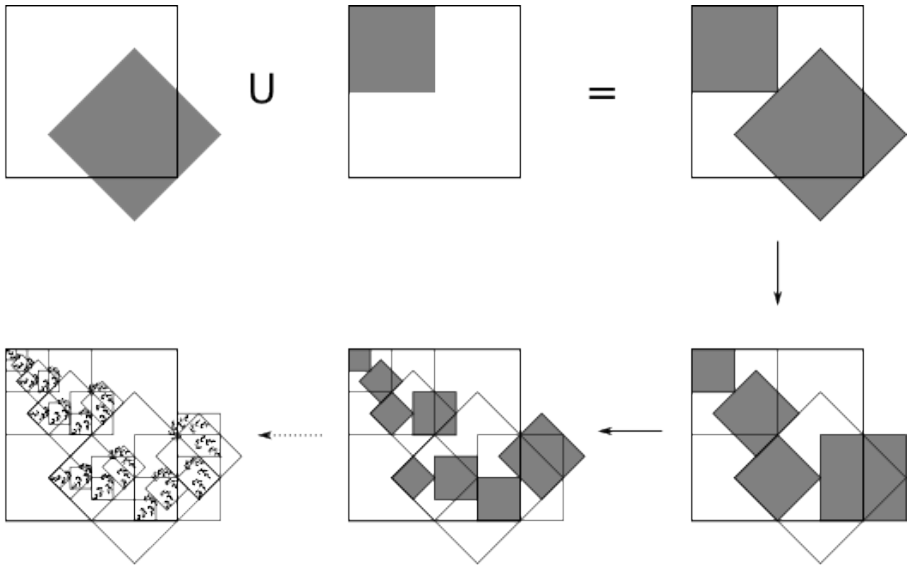


Figure 8.5: An IFS Example.
This is a file from the Wikimedia Commons, `tinyurl.com/dbt9uaxy`

known examples are a simple sequence of discrete values, and the (continuous) solutions of a system of differential equations. Here are examples.

• An infinite mapping of the form $x_{n+1} \leftarrow f(x_n)$. Such a limit may exist for a wide range of initial values x_0, but may be very sensitive to those values. Even the slightest change in x_0 may result in a completely different behavior of the sequence elements x_i, and may lead to different limit. A well-known example of such mapping is the logistic map, reference [logistic map 24]. This is the simple nonlinear mapping $x_{n+1} \leftarrow r\, x_n(1 - x_n)$, whose behavior is very sensitive to the value of r and to the initial x_0. For some values, the sequence diverges to negative infinity, for other values it may either converge to a unique value (a single attractor) or it may oscillate between two values of x.

• The Lorenz attractor. The story of Edward Lorenz and his discovery of the behavior of nonlinear processes and chaotic systems is told on Page 356.

The Lorenz system (or Lorenz attractor) is a system of ordinary differential equations first studied by Lorenz. His first system consisted of the equations

$$\frac{dx}{dt} = \sigma(y - x), \quad \frac{dy}{dt} = x(\rho - z) - y, \quad \frac{dz}{dt} = xy - \beta z.$$

For certain parameter values and initial conditions, this system exhibits chaotic solutions. Even a slight change in one of the parameters or conditions leads to a widely different and unpredictable solution. This behavior has resulted in the term "chaos" or "chaotic behavior" applied to such systems.

It is perhaps useful to think of such a system as being globally stable—because it has

solutions to a wide range of parameter values and initial conditions—but also locally unstable, because the solutions are very sensitive to the initial conditions, making it impossible to predict the solution after even a slight change in the conditions.

■ L-systems. An L-system, reference [L-systems 24], is a type of formal grammar. You start with a single symbol (the axiom) drawn from some alphabet of symbols. You then replace that symbol with a substring of symbols according to some production rule. You continue to replace each of the elements of the resulting substring with another substring according to other production rules. This iterative process may result in a vast, complex string of symbols that is meaningful because it may describe the growth of a plant or a bacteria, it may describe the morphology (shape) and the growth pattern of a complex plant, and it may also result in a infinitely-complex mathematical pattern that can be considered a fractal.

L-systems grew from the ideas, in 1968, of the biologist and botanist Aristid Lindenmayer, who used them to describe the behavior of plant cells and to model the growth processes of plant development.

The formal definition of an L-system is a triplet $G = (V, \omega, P)$, where V is the alphabet of symbols. It consists of variables (symbols that can be replaced) and terminals (constants that cannot be replaced). The initial value (or axiom) ω is a symbol or a string of symbols from V. P stands for the set of production rules which describe how to replace symbols in the string that is being constructed. A rule consists of two strings, the predecessor (which is being replaced by the rule) and the successor (its replacement). A symbol X which is in the alphabet but which is not the predecessor of any rule, is called a terminal. It is never replaced.

The rules are applied simultaneously to all the symbols of the current string, as the following example illustrates. Consider the L-system defined by $V = (A, B)$, $\omega = A$, and $P\colon (A \to AB), (B \to A)$. Then successive generations produce the following substrings (where the vertical bars are for easier reading):

0: A
1: AB
2: $AB|A$
3: $AB|A|AB$
4: $AB|A|AB|AB|A$
5: $AB|A|AB|AB|A|AB|A|AB$
6: $AB|A|AB|AB|A|AB|A|AB|AB|A|AB|AB|A$
7: $AB|A|AB|AB|A|AB|A|AB|AB|A|AB|AB|A|AB|A|AB|AB|A|AB|A|AB$

Notice that the strings produced by this system have lengths that are Fibonacci numbers. Also the ratio of As to Bs converges to the golden ratio.

◇ **Exercise 8.1:** (A joke.) Notice that these lengths start with 1, 2, 3, and 5, whereas the Fibonacci sequence starts with 1, 1, 2, 3, and 5. Is this because of the axiom of choice?

The next example of an L-system creates a fractal binary tree. It is defined by an alphabet of two variables 0 and 1 and two constants "[" and "]". The axiom is 0 and the rules are $(1 \to 11)$ and $(0 \to 1[0]0)$. Just three iterations of this rule produce the long complex strings

1: 1[0]0

2: 11[1[0]0]1[0]0
3: 1111[11[1[0]0]1[0]0]11[1[0]0]1[0]0

In order to actually draw a fractal binary tree, we employ two concepts that are familiar to any computer programmer. They are a LIFO (last in, first out) stack and turtle graphics. The LIFO principle is to pop out of the stack the last item pushed into it. Turtle graphics is a basic method for drawing digital graphics. It is based on an imaginary turtle that can move in (two- or three-dimensional) space and has at any time a location, an orientation, and a pen that can be up or down (other attributes, such as color and line thickness, can also be assigned to the turtle).

A complete drawing is generated by specifying a string of triplets, each consisting of (1) move/turn (2) how much, and (3) pen up/down. Typical examples are (forward, 10 steps, pen up), (turn left, 45°, pen down), (back, 5 steps, pen down).

Based the concept of turtle graphics, the tree is drawn by the following rules:

- 0: draw a line segment ending in a leaf
- 1: draw a line segment
- [: push position and angle into the stack, turn left 45°
-]: pop position and angle out of the stack, turn right 45°

The iterations above, combined with the tree drawing rules, produce the tree shown growing in Figure 8.6

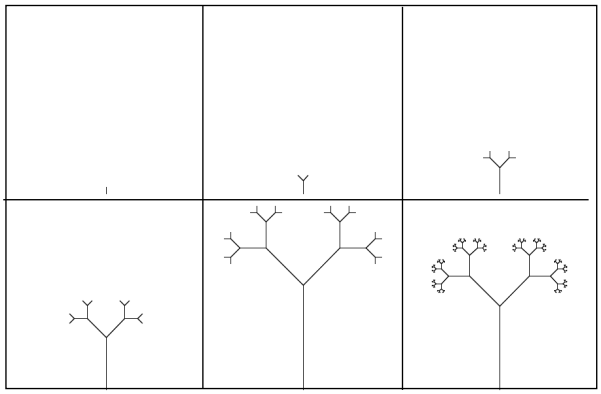


Figure 8.6: Constructing a Fractal Binary Tree.

◇ **Exercise 8.2:** Implement a recursive function to compute and display a fractal binary tree (aka as a symmetric binary trees. Use any programming language.

■ Escape-time fractals. In this approach to fractal design, an area R in the complex plane is chosen. A point c in R is chosen, and a nonlinear iteration is performed with c as its seed, to generate a sequence of points (complex numbers). If the sequence diverges, point c is ignored, but if the sequence is bounded (it converges to a point or it oscillates within a finite region of the complex plane), point c is added to the fractal being generated. The process is repeated for every point in R (well, for as many points as practically possible).

The first fractal generated and researched by this approach is the Mandelbrot set, described here, but many other escape-time fractals are known and have been researched and described in detail, among them the Julia set, burning ship fractal, nova fractal, and Lyapunov fractal.

The Mandelbrot set is defined by the seed $z = 0$ and the iteration $z_{i+1} = z_i^2 + c$ for an arbitrary complex number c. The iteration is performed on as many numbers c as possible, and those values of c for which the iteration remains bounded are appended to the set. The set was envisioned and (roughly) drawn by others, but it was Mandelbrot who popularized it and thereby proved the little-known adage that an invention is often named after the *last* person who invented it.

It is trivial to see that for $c = 1$, the iteration runs through 0, 1, 2, 5, 26, ... and diverges to infinity, but for $c = -1$, the recursive sequence 0, -1, 0, -1, 0, ..., is bounded, which is why -1 belongs to the Mandelbrot set. The shape of this set is not shown here, because it is very familiar and it appears in countless places.

The burning ship fractal is similar to the Mandelbrot set, with the difference that instead of squaring z_i, each iteration squares the sum of the absolute values of the real and imaginary parts of z, Thus

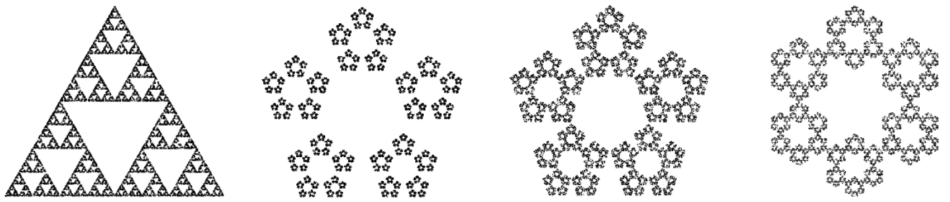
$$z_{n+1} = (|Re(z_n)| + i|Im(z_n)|)^2 + c, \quad z_0 = 0.$$

■ The Chaos game. This fun and unexpected approach to fractal design is based on randomness. Start with a regular n-gon and choose a point $\mathbf{S}$ (a seed) inside it at random. Start a loop where a new point is computed in each iteration from its predecessor. Starting with $\mathbf{S}$, a polygon vertex $\mathbf{v}$ is chosen at random, and a new point $\mathbf{Q}$ is generated along the segment connecting $\mathbf{S}$ to $\mathbf{v}$. If the length of this segment is denoted by L, then $\mathbf{Q}$ is placed at a distance of $r \cdot L$ from $\mathbf{S}$, where r is often, but not always, less than 1. The next iteration generated a point $\mathbf{P}$ from $\mathbf{Q}$ in the same way. If a noticeable, non-random pattern emerges after many iterations, the final figure can be "cleaned" by removing some of the initial points, which tend to be outliers.

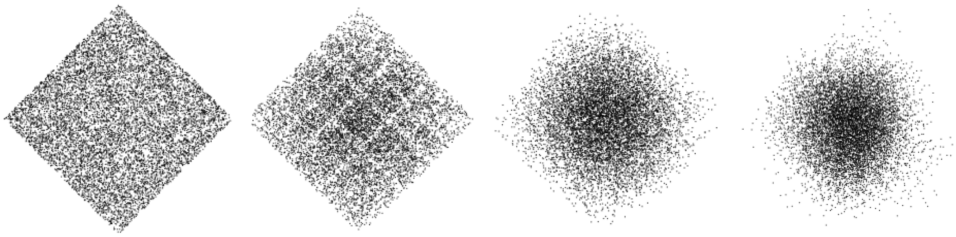
This game is fully defined by the values of n and r. With a bit of luck, a regular pattern starts to emerge after many iterations, but if it happens that no discernible pattern appears, the game can be repeated with different parameters or rules. Here are some variations that can produce well-defined and attractive fractals.

1. The next vertex cannot be the same as the current vertex.
2. The next vertex must be one place away (anti-clockwise) from the current vertex.
3. The next vertex cannot be a near neighbor of the current vertex if the current vertex and its predecessor are the same.

The top row of Figure 8.7 illustrates obvious fractals that correspond to various values of n and r. The bottom row of the figure shows how $n = 4$ (a square) resists the chaos game for several values of r, although $r = 0.6$ promises an eventual non-random pattern. The patterns in the figure were computed and displayed by the Mathematica notebook `ChaosGame.nb`, available from reference [ChaosGameNotebk 24]. Wikipedia article `Chaos_game` has more examples. The popular mathematical software `Geogebra` can also be used to produce Chaos Game fractals.



n=3, r=1/2 n=5, r=1/3 n=5, r=3/8 n=6, r=1/3



n=4, r=0.5 n=4, r=0.6 n=4, r=0.75 n=4, r=0.9

Figure 8.7: Fractals and Non-Fractals Generated by the Chaos Game.

8.1.1 Iterated Function Systems (IFS)

Figure 8.8 illustrates six IFS fractals whose Mathematica codes are listed in Figure 8.9. The following paragraphs explain the process that creates the simple fractal known as the Sierpinski triangle. Start with an arbitrary triangle T. (The figures use an equilateral triangle, which generates a symmetric Sierpinski triangle.) Choose a point p, inside or outside the triangle, at random, and repeat the following process to generate the set of points that would form the Sierpinski triangle.

1. Choose one of the three triangle vertices q at random.
2. Compute the arithmetic average of the coordinates of p and q.
3. Generate a new point s at the average. This point is now at half the distance of p from q.
4. Set p to s.

When this simple process is repeated, choosing vertices q at random, more and more points are generated inside the original triangle T (a few of the initial points may be outside T) and they start resembling the Sierpinski triangle of Figure 8.4.

This process is dynamic, it is a collection of activities or tasks without a predetermined sequence of execution. Because of the use of randomness, this process is also nondeterministic. However, regardless of the sequence of random decisions made during this process, the result always looks the same. The discipline of IFS (iterated function systems) is therefore an example of an application of mathematics for using randomness to create a regular, non-random object. We also know, from experience with chaotic processes, that a deterministic process can also end up with an object that looks random. We can therefore conclude that mathematics "can have it both ways," and has very wide applications.

Here are the details of the Mathematica code for the Sierpinski triangle (the first listing of Figure 8.9).

In forest folds where ferns abide,
A secret form begins to hide
A curve, a crack, a curling seam,
That echoes deep a patterned dream.

From snowflake stars to coral's spine,
Each twist repeats, each edge aligns.
A branch becomes a mirrored twig,
A mountain's jag is just a jig.

Zoom closer still, what do you see?
The same old shape repeats with glee!
Self-similar, yet ever new,
It plays with scale and puzzles you.

No compass drew this tangled lace,
No ruler dared to map its space.
Yet nature hums in fractal tones,
In broccoli heads and lightning's bones.

A dance of math in looping play,
Infinity on bold display.
It's art, it's code, it's life unfurled
A fractal breathes throughout the world.

Whispers of the Infinite.

A triangle stands, so bold, so still,
Three corners cast with patient will.
But watch it closely, don't look away,
It's not just shape, it's math at play.

A central bite, a hollow trace,
Three more arise to take its place.
Then each of them begins to fade,
In perfect holes, more holes are made.

A puzzle drawn with steady hand,
That folds itself like desert sand.
No curves to charm, no swooping arcs,
Just silent steps and shrinking marks.

It's simple rules that hide such depth,
A breath of space with every breath.
From one to three, from three to nine,
A lace of voids in perfect line.

A fractal born from rule and grace,
A triangle laced with empty space.
It whispers low: "Infinity
Can fit within the likes of me."

Sierpinski's Secret.

Poems by chatGPT 4.o, in the style of Theoni Pappas.

The Mathematica code of Figure 8.9 may look opaque to non-experts, so here is a detailed description of it for the Sierpinski triangle, the first fractal in the figure. Variable `n` is the number of points, which should be in the thousands for a sharp figure. Variable `lc` is a list of pairs of coordinates for the points that will be computed later. It starts with one pair, which can be anything. Similarly, `nc` is an empty list into which the pairs of `lc` are eventually transferred to become Mathematica points (in Mathematica there is a difference between a pair of coordinates and a point). List `p` contains the coordinates of the three triangle vertices.

The first `Do` loop places a random vertex number (between 1 and 3) in `r`, computes the average coordinates of the chosen vertex and the current pair in `lc`, and appends it to `lc`. The next `Do` transfers the pairs of coordinates, one by one, from `lc` to become points in `nc`. The `graphics` command display the points with the original triangle from list `p`.

◇ **Exercise 8.3:** Read about the Sierpinski carpet (or gasket). In a programming language of your choice, implement a recursive function to compute and display it.

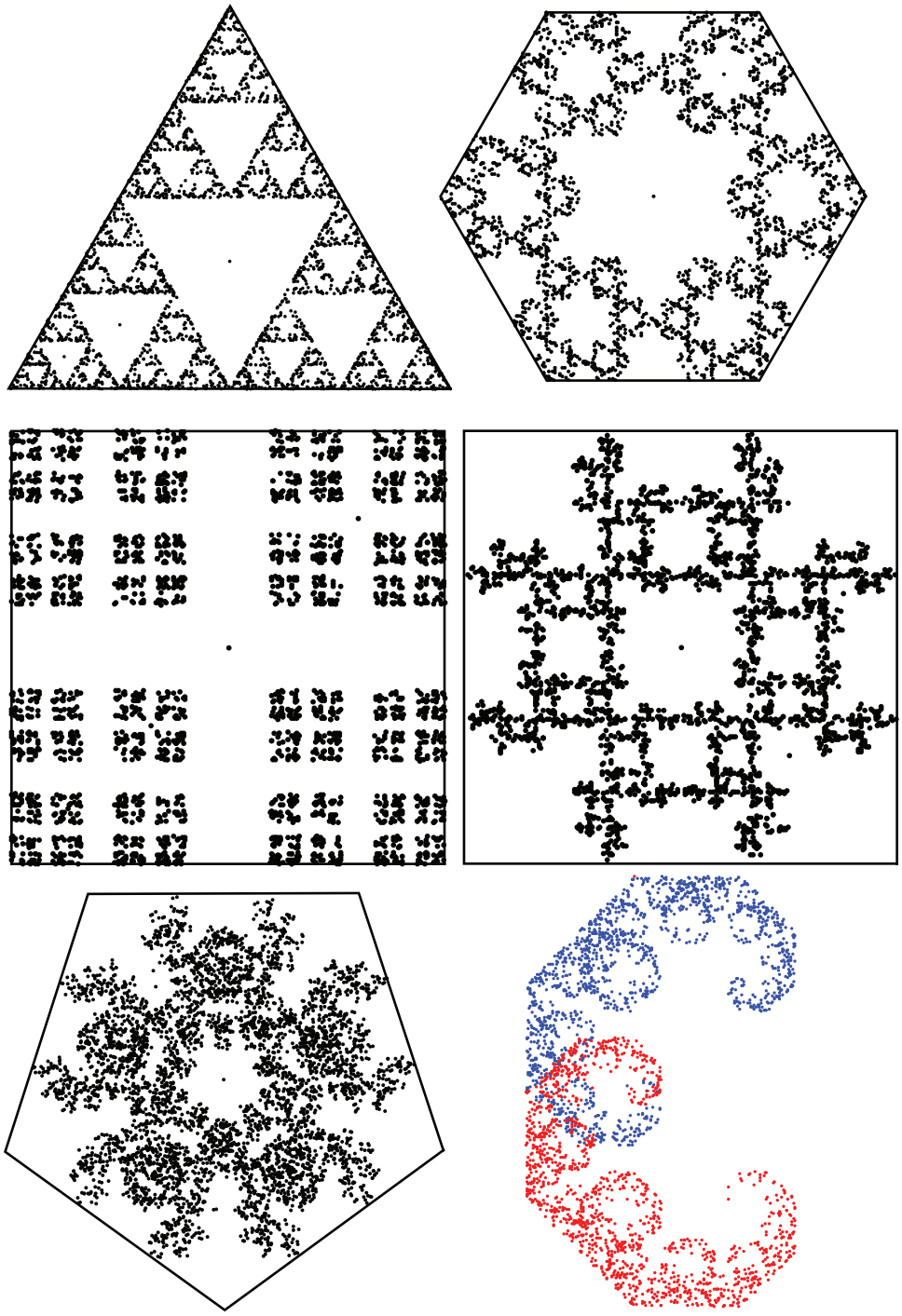


Figure 8.8: Six IFS Fractals.

```
(*Sierpinski triangle in IFS. The next point is placed half the distance of the
 current point from a random vertex v*)
n=50; lc={{0,0}}; nc={};
p={{1,0},{-1/2.,Sqrt[3.]/2},{-1/2.,-Sqrt[3.]/2}};
Do[{r=RandomInteger[{1,3}], ne=(lc[[i]]+p[[r]])/2,AppendTo[lc,ne]},{i,n}]
Do[AppendTo[nc,Point[lc[[i]]]],{i,n}]
Line[{p[[1]],p[[2]],p[[3]]}]
Graphics[{nc,Line[{p[[1]],p[[2]],p[[3]],p[[1]]}]}]

(*Sierpinski hexagon, scale points 2/3 of the way toward a vertex*)
n=3000; lc={{0,0}}; nc={};
p={{1.,0},{0.5,Sqrt[3.]/2},{-0.5,Sqrt[3.]/2},{-1.,0},
 {-0.5,-Sqrt[3.]/2},{0.5,-Sqrt[3.]/2}};
Do[{r=RandomInteger[{1,6}], ne=(lc[[i]]+2p[[r]])/3,AppendTo[lc,ne]},{i,n}]
Do[AppendTo[nc,Point[lc[[i]]]],{i,n}]
Graphics[{nc,Line[{p[[1]],p[[2]],p[[3]],p[[4]],p[[5]],p[[6]],p[[1]]}]}]

(*Sierpinski square, scale points 60\% of the way toward a vertex*)
n=3000; lc={{0,0}}; nc={}; p={{1.,0}, {0,1.}, {-1.,0}, {0,-1}};
Do[{r = RandomInteger[{1, 4}], ne = (lc[[i]] + 1.5 p[[r]])/2.5,
  AppendTo[lc, ne]}, {i, n}]
Do[AppendTo[nc, Point[lc[[i]]]], {i, n}]
Graphics[{nc, Line[{p[[1]], p[[2]], p[[3]], p[[4]], p[[1]]}]}]

(*Constrained Sierpinski square. Move points half the way toward a \
vertex but don't choose same vertex twice in a row*)
n = 3000; lc = {{0, 0}}; nc = {};
p = {{1., 0}, {0, 1.}, {-1., 0}, {0, -1}};
r = ri = 0;
Do[{Until[r != ri, r = RandomInteger[{1, 4}]], ri = r,
   ne = (lc[[i]] + p[[r]])/2, AppendTo[lc, ne]}, {i, n}]
Do[AppendTo[nc, Point[lc[[i]]]], {i, n}]
Graphics[{nc, Line[{p[[1]], p[[2]], p[[3]], p[[4]], p[[1]]}]}]

(*Constrained chaos game on pentagon. Move points half the way toward
a vertex but don't choose same vertex twice in a row*)
n = 5000; lc = {{.1, .1}}; nc = {}; r = ri = 0;
p = {{1,0}, {0.309,0.951}, {-0.809,0.588}, {-0.809,-0.588}, {0.309,-0.951}};
Do[{Until[r != ri, r = RandomInteger[{1, 5}]], ri = r,
  ne = (lc[[i]] + p[[r]])/2, AppendTo[lc, ne]}, {i, n}]
Do[AppendTo[nc, Point[lc[[i]]]], {i, n}]
Graphics[{nc,
  Line[{p[[1]], p[[2]], p[[3]], p[[4]], p[[5]], p[[1]]}]}]

(*Colored Levy curve in IFS*)
n = 3000; lc = {{.1, .1}}; clr = {{1}}; bp = {}; rp = {}; pt = {1,1};
A = {{.5, .5}, {-.5, .5}};
B = {{.5, -.5}, {.5, .5}};
f[1, p_] := A . p;
f[2, p_] := B . p - {.5, .5};
Do[{r = RandomInteger[{1, 2}], AppendTo[clr, {r}],
AppendTo[lc,f[r,pt]], pt=lc[[i]]}, {i,n}] (*Append pairs from f[r, ] to lc*)
Do[If[clr[[i]] == {1}, AppendTo[bp, Point[lc[[i]]]],
  AppendTo[rp, Point[lc[[i]]]]], {i, n}]
(*Transfer pairs from lc to become points in bp and rp*)
Graphics[{{Blue, bp}, {Red, rp}}] (*display bp and rp in blue and red*)
```

Figure 8.9: Six IFS Fractals.

IFS with affine transformations.

The last Mathematica code of Figure 8.9 is different from the rest. It defines two matrices and two functions `f1[x]` and `f2[x]`. Each function computes the (x, y) pair of coordinates of the next point of the fractal, and these coordinates are later appended to either list `bp` (blue point) or list `rp` (red point) depending on the color of the point in the final fractal. The color depends on which function, `f1` or `f2`, was chosen at random to compute the point. Color information for each newly computes point is appended to list `clr`.

The important feature of this particular code is the use of transformation matrices in generating the individual points that make the fractal. In general, the next point should depend on the current point, as well as on other data such as the coordinates of a fixed point or some user-defined parameters. In order to speed up the many computations needed for an IFS fractal, the next point should be a linear transformation of the current point, and such transformations are referred to as *affine*.

From Wikipedia: In Euclidean geometry, an affine transformation or affinity (from the Latin, affinis, "connected with") is a geometric transformation that preserves lines and parallelism, but not necessarily Euclidean distances and angles.

Section 3.6 discusses two-dimensional transformations, which are affine, and those useful maps are presented here in terms of two small matrices, one for translations and the other one for all the other transformations. In general an affine two-dimensional transformation maps a pair (x, y) to the pair (x', y') defined by the linear operations $(x, y) \to (x', y') = (ax + by + \alpha, cx + dy + \beta)$. This can also be expressed in matrix notation as

$$(x', y') = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

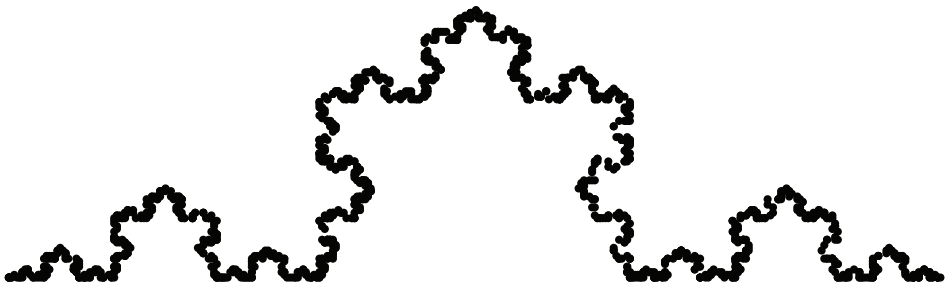
Figure 8.10 illustrates how the six quantities a, b, c, d, α, and β cause certain two-dimensional transformations, as well as the fact that affine transformations are the sum of (1) a translation, caused by a 2×1 matrix whose elements t_x and t_y may both be zeros, and (2) a 2×2 matrix whose elements are responsible for rotations (clockwise or counterclockwise), reflection (about the x axis, the y axis, or both), scaling (zooming or shrinking in the x or y directions or both), and shearing (again, in either or both directions).

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad \begin{bmatrix} c_x & 0 \\ 0 & c_y \end{bmatrix} \quad \begin{bmatrix} 1 & c_x \\ c_y & 1 \end{bmatrix} \quad \begin{bmatrix} t_x \\ t_y \end{bmatrix}.$$

Identity x-reflection CC rotation scaling shearing translation

Figure 8.10: Affine Transformations.

As a result, a general IFS algorithm employs a set of affine transformations that are chosen randomly and are applied many times to generate the list of points that make up the final fractal figure. The following example is typical. It includes four affine transformations to create an IFS image of the well-known Koch curve, described and illustrated (as Koch snowflake) in Figure 8.3.



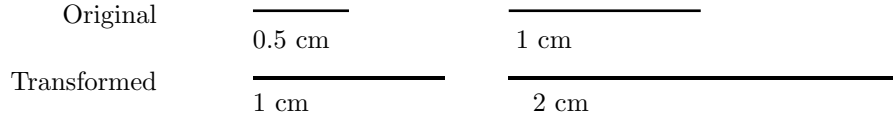
```
n=1500;lc={{0,0}};nc={};
sc=1/3.; s6=Sin[60 Degree]/3.;
M1={{sc,0},{0,sc}}; (*Scale by sc*)
M2={{sc/2,-s6},{s6,sc/2}}; (*Scale and rotate CC 60 D*)
M3={{sc/2,s6},{-s6,sc/2}}; (*Scale and rotate C 60 D*)
f[1,p_]:=M1 .p;
f[2,p_]:=M2 .p+{sc,0}; (*+Translate right*)
f[3,p_]:=M3.p+{0.5,s6}; (*+Translate up and right*)
f[4,p_]:=M1 .p+{2sc,0};(*+Translate right by 2sc*)
Do[{r=RandomInteger[{1,4}],ne=f[r,lc[[i]]],AppendTo[lc,ne]},{i,n}]
Do[AppendTo[nc,Point[lc[[i]]]],{i,n}]
Graphics[nc]
```

Figure 8.11: The Koch Snowflake in IFS.

◇ **Exercise 8.4:** Implement The Koch Snowflake in a programming language of your choice.

The following geometric properties of a figure are preserved by affine transformations:

- Collinearity. A set of collinear points remains collinear after an affine transformation.
- Parallelism. A set of parallel lines remains parallel.
- Set convexity. A set of points is convex if every chord through the set remains entirely inside the set. Even more. When a convex set is affinely transformed, its extreme points are transformed such that they become the extreme points of the new set.
- Lengths of parallel lines are transformed in the same way as other objects. More precisely, the ratios of parallel line segments are preserved, as illustrated in the figure.



8.2 Weierstrass Function

One of the central concepts in mathematics is that of a function. The formal definition of a function is straightforward. A function f from a set X to a set Y is an assignment of one element of Y to each element of X. The set X is called the domain of the function and the set Y is called the codomain of the function (in the past the codomain was referred to as the range of the function). The notation $f : X \to Y$ is used to indicate a function and its two sets, while the familiar notation $y = f(x)$ implies that set element x is mapped to set element y by function f.

A continuous function is one that satisfies the following property: a small variation of its argument induces a small variation of the value of the function. This implies that there are no abrupt changes—discontinuities, jumps, or holes—in its value. The converse is also true. If we see a small change in the value of a function, we can trust that it was caused by a small change in its argument.

A discontinuous function is a function that is not continuous. The following are simple discontinuous functions:

- The simple function $y = 1/x$ is discontinuous at $x = 0$.
- The Heaviside step function $H(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$
- The sign function, signum $\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$
- The function $\sin(1/x^2)$ is discontinuous at $x = 0$. It also oscillates infinitely many times as its argument approaches zero (either from below or from above).

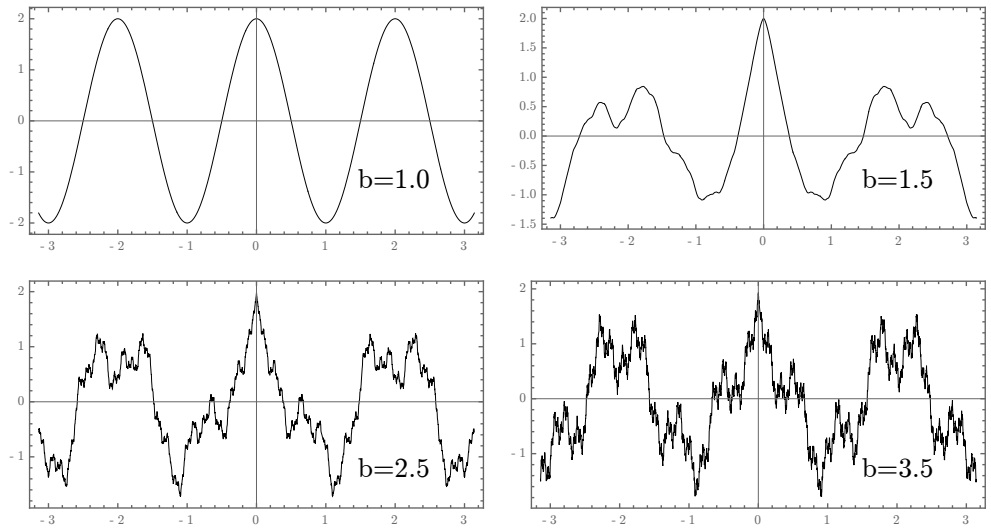
A function is differentiable if its derivative exists at every point in its domain. The graph of a differentiable function is never vertical; it does not have points where the slope is infinite. A differentiable function has no breaks, angles, or cusps, and is smooth in the sense that it is locally well approximated as a linear function at each interior point.

Many functions have one or several points where they are nondifferentiable; they have a point or a cusp, but may still be continuous at those points. Examples are $y = |x|$ and $y = \sqrt{|x|} + 1/2$. The function $|x^2 - 1|$ is nondifferentiable but continuous at the two points $x = \pm 1$. A similar example is $|x - 1| + |x + 1|$. In contrast, the step function $y = \lfloor x \rfloor$ and the trigonometric function $\tan(x)$, even though generally smooth, are nondifferentiable at infinitely many points, and are also discontinuous at those points.

◇ **Exercise 8.5:** Find examples of simple functions that are nondifferentiable at infinitely many points.

In the past, mathematicians had the notion that a continuous function could be nondifferentiable on at most a set of isolated points, but this belief was shattered in 1872, when Karl Weierstrass managed to artificially construct a function, now named after him, that is everywhere continuous but nowhere differentiable. It is as if this function changes direction every time it proceeds from a real number to its successor, except that real numbers don't have successors (see Page 99). Figure 8.12 gives an

idea of the shape of the graph of this function, which obviously cannot be viewed in full detail. Even its general shape was unknown and had to wait for the power of the modern computer. When we try to zoom in on a small region of this function, it looks jagged at any magnification. Any small parts of it are irregular, spiky, and serrated.




```
a = 0.5; b = 3.5;
f[x_] := Sum[a^n Cos[b^n Pi x], {n, 0, 100}];
Plot[f[x], {x,-Pi,Pi}, Frame->True, AspectRatio->0.5, PlotStyle->{Black, Thin}]
```

Figure 8.12: Weierstrass Function for $a = 0.5$ and Various b.

Such a function cannot be expressed simply in terms of basic arithmetic operations and has the form of an infinite sum of cosines. It is therefore a special case of a Fourier series (an expansion of a periodic function into a sum of trigonometric functions). The original function presented by Weierstrass was the infinite sum

$$\frac{\cos(3\pi x)}{2} + \frac{\cos(3^2\pi x)}{2^2} + \frac{\cos(3^3\pi x)}{2^3} + \cdots,$$

but the modern definition is

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x), \text{ where } a \in (0,1), b \text{ is a positive odd integer, and } ab < 1 + \frac{3}{2}\pi.$$

Weierstrass's contemporaries deplored his discovery. They wre shocked and offended by it and referred to it as pathological, a monster, "lamentable scourge," and "an outrage against common sense." All of which are understandable, and yet this function, as well as similar "cousins," discovered at about the same time, refused to go away. It took mathematicians decades to come to terms with this monster which trampled over many

of their old, well-established theories and proofs. It took time for people to realize that the new monster functions provide a key that opens new fields and concepts in mathematics, and they should be acceptable as part of this branch of knowledge. Today, we simply refer to them as examples of fractals.

> I recoil with fear and loathing from that deplorable evil, continuous functions with no derivative.
>
> —Charles Hermite (1893), on the Weierstrass function .

Reference [Kucharski 14] tells the history of the Weierstrass function and the beginning of the acceptance of fractals.

A bit of history. Already in 1861, Bernhard Riemann claimed that the function

$$f_a(x) = \sum_{k=0}^{\infty} \frac{\sin(k^a \pi x)}{k^a \pi},$$

is nondifferentiable on a dense set of the real numbers, but he did not provide a proof. Reference [Weisstein 24] shows that for rational arguments $x = p/q$, this function can be computed as the *finite* sum

$$f\left[\frac{p}{q}\right] = \frac{\pi}{4q^2} \sum_{k=1}^{q-1} \frac{\sin\left[\frac{k^2 p\pi}{q}\right]}{\sin^2\left[\frac{k\pi}{2q}\right]}.$$

8.3 Continuity

Continuity is one of the key concepts in the study and analysis of real and complex functions.

What does it mean to say that a function is continuous or that it is continuous in a certain interval? Here is some explanations, arranged roughly in increasing order of complexity.

A curve is continuous if it can be drawn without lifting the pen off the paper.

It is continuous if it doesn't have "holes" or "jumps."

Continuity is the opposite of discreteness.

A function is continuous at an input if that output is what we intuitively expect it to be by examining nearby outputs.

The term continuous means that the value of $f(x)$ at a point p equals the limit of values as the inputs approach p.

A function $f(x)$ is continuous at point p if $\lim_{x\to p} f(x) = f(p)$.

It is continuous if a small variation of the argument induces a small variation of the value of f.

Alternatively, if $f(x)$ gets closer to $f(p)$ as x gets closer to p.

A function is continuous if arbitrarily small changes in its value can be assured by restricting to sufficiently small changes of its argument.

An infinitesimal change in x necessarily produces an infinitesimal change in y (Cauchy, 1821).

The limit of a function $f(x)$ as x approaches p is a number L with the following property: given any target distance from L, there is a distance from p within which the values of $f(x)$ remain within the target distance.

A definition based on the topological concept of a neighborhood. A function f from a domain X to a codomain (range) Y is continuous at a point $x \in X$ if and only if for any neighborhood V of $f(x)$ in Y, there is a neighborhood U of x such that $f(U) \subseteq V$.

Finally, the formal, rigorous (ε, λ) definition of continuity at a point is: A function $f(x)$ is continuous at point p if for any real positive ε there exists a real positive λ such that $|x - p| < \lambda$ implies $|f(x) - f(p)| < \varepsilon$.

Today, we believe that the definition of continuity was first introduced into mathematics in 1817 by Bernard Bolzano. In 1861, it was used by Karl Weierstrass and popularized by him and others. We still use this form today, with the only difference, introduced by Hardy in 1908, that the arrow is placed under the limit symbol instead of following it.

Because of the infinite number of functions and their varieties, there may be unusual, strange cases of one-sided limits that are different. Consider the limits $\lim_{x\to p^+} f(x)$ and $\lim_{x\to p^-} f(x)$. In the former, the limit is computed by starting with $x > p$ and decreasing x toward p. In the latter case, x starts smaller than p and is increased. We say that the limit exists if the two limits above are equal.

Simple examples of one-sided limits are:

1. Function $y(x) = |x|/x$ at $x = 0$. The limit $\lim_{x\to 0} y(x)$ is one-sided. When zero is approached from the right, the limit of y is 1, but when zero is approached from the left, the limit is -1. Hence, y is discontinuous at zero.

2. Function $y(x) = 1/(x + 1)$ at $x = 0$. The limit $\lim_{x\to 0} y(x)$ is one-sided. When zero is approached from the right, the limit of y is ∞, and when zero is approached from the left, the limit is $-\infty$. Hence, this function is also discontinuous at zero.

Earlier in this Chapter we discuss functions that are everywhere continuous but nowhere differentiable, as well as functions that are everywhere discontinuous. In this section we are interested in functions that are continuous at one point only or at a finite number of points, or at an infinity of points. Such functions used to be referred to as pathological, weird, or strange, but today we realize that they are of interest because they provide deeper understanding of the mathematical concept of continuity.

Intuitively, a function that is everywhere discontinuous jumps up or down at every point in its domain, but intuition, even though useful in mathematics, must always be checked and rephrased carefully, so here is a more rigorous definition of function discontinuity. A function f is nowhere continuous if for each point x in its domain there is some $\varepsilon > 0$ such that for every $\delta > 0$ we can find a point y such that $|x - y| < \delta$ but $|f(x) - f(y)| \geq \varepsilon$. In other words, no matter how close f gets to any fixed point x, there are points closer to x at which the function takes values that are not close to its value at x.

We expect an everywhere-discontinuous function, like an everywhere nondifferentiable function, to be more complex than a polynomial, a logarithm, or a trigonometric function, and this is generally, but not always, true. We continue with a few examples of these functions, which are sometimes referred to as pathological.

- Dirichlet function, named after Johann Dirichlet. Its definition is

$$f(x) = \begin{cases} 1, & \text{if } x \in Q, \\ 0, & \text{if } x \notin Q. \end{cases}$$

Both the rationals and irrationals are dense sets, and each is dense in the other. Therefore, the best that can be done graphically is to draw each case as a horizontal set of dense dots; it is impossible to get a better drawing.

This function is discontinuous at any real point, rational or irrational, because any rational point (where $f(x) = 1$) is surrounded by infinitely many irrational points (where $f(x) = 0$). The function oscillates infinitely many times even at the shortest possible distance from any point on the real number line. Consider the case $f(8)$. At this point, Dirichlet's function is 1, but as we get closer to 8, either from the left or from the right, we pass through infinitely many irrational numbers, where the function is zero; not close to 1.

We can find sequences of rational numbers that converge to 8, such as $a_n = 8 + 1/n$. The sequence $f(a_n)$ converges to 1, because it is always 1, but the sequence $f(8 + \pi/n)$ is always irrational and therefore all zeros and it converges to zero at 8.

We conclude that it is impossible to find a general limit for $f(x)$ at any point, because different sequences x_n that converge to 8 would converge to different limits of $f(x_n)$. This justifies the claim that such a limit does not exist and proves that Dirichlet's function is everywhere discontinuous.

The following proof that the Dirichlet function is everywhere discontinuous is quoted from reference [WikiDirichlet 24]:

If y is rational, then $f(y) = 1$. To show that Dirichlet function is not continuous at y, we need to find an ε such that no matter how small we choose λ, there will be points z within λ of y such that $f(z)$ is not within ε of $f(y) = 1$. In fact, $1/2$ is such an ε. Because the irrational numbers are dense in the reals, no matter what λ we choose, we can always find an irrational z within λ of y, and $f(z) = 0$ is at least $1/2$ away from 1.

If y is irrational, then $f(y) = 0$. Again, we can take $\varepsilon = 1/2$, and this time, because the rational numbers are dense in the reals, we can pick z to be a rational number as close to y as is required. Again, $f(z) = 1$ is more than $1/2$ away from $f(y) = 0$.

- The Dirichlet function brings us to an even more unusual function that is somewhat similar to Dirichlet's. It is not smooth, it is infinitely fragmented, but there is one point, namely zero, where it is continuous. Reference [Cummings 19] refers to this strange creature as a modified Dirichlet's function. It is defined as

$$f(x) = \begin{cases} x, & \text{if } x \in Q, \\ 0, & \text{if } x \notin Q. \end{cases}$$

Thus, over the rationals, this function is a fragmented straight line with a 45° slope, while on the irrationals, it is the fragmented x-axis. The previous analysis of the original Dirichlet function implies that the modified version is discontinuous both on the x-axis and on the diagonal line, but what about the point $x = 0$?

It turns out that all sequences a_i that converge to zero result in sequences $f(a_i)$ that converge to $f(0)$. As we move over the real numbers toward zero, the function

oscillates between zero (when the argument is irrational) and x (when the argument is rational), but the oscillations dampen as we approach zero, and the limit of $f(a_i)$ is zero, declaring the real number 0 a point where our modified, weird function is continuous.

- The Thomae function, sometimes called the popcorn function, and named after Carl Thomae. This is another modification of the Dirichlet function.

$$h(x) = \begin{cases} 1, & \text{if } x = 0, \\ \frac{1}{q} & \text{if } x \in Q \text{ and } x = \frac{p}{q}, \text{ where the integer } p \text{ and the natural } q \text{ are coprime,} \\ 0 & \text{if } x \notin Q. \end{cases}$$

The term coprime (or relatively prime) means that 1 is the only positive integer that divides both p and q. The Thomae function is well defined, because every rational number can be uniquely represented with an integer p and a natural q which are coprime.

It is easy to convince ourselves that at every rational point r, $h(r)$ is positive, but as we move on the number line toward point r, we go through infinitely many irrational numbers, where $h(x)$ is zero. This function is therefore discontinuous at every rational point.

Now choose an irrational number c. Given an arbitrary positive ε, we can always find a real N such that $\frac{1}{N} < \varepsilon$. There must therefore be a region around c where all x values satisfy $h(x) < \frac{1}{N} < \varepsilon$. This is true, because within any given distance of c there are only a finite number of rationals which have a denominator of N or smaller. Find the one that is nearest c, call it d, and choose $\delta = |c - d|$. Now, all the rationals in the open interval $(c\delta, c+\delta)$ have denominators that are larger than N, implying that all $h()$ values within this interval are smaller than $1/N$.

The unusual Thomae function is therefore continuous at every irrational number and discontinuous at every rational number

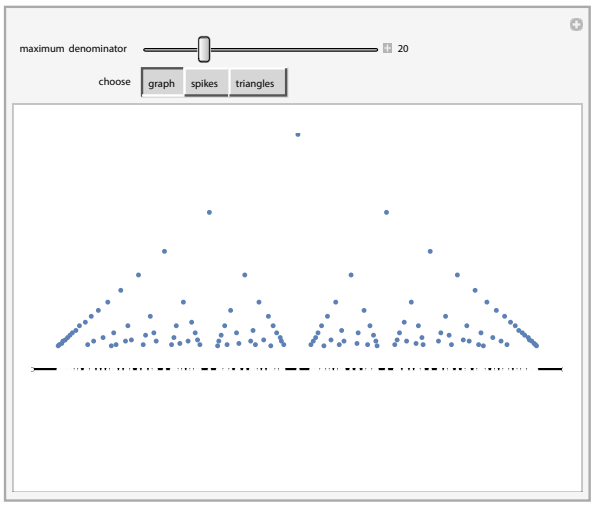
The definition of Thomae function is deceptively simple, but drawing an approximate graph of it requires some thinking. It requires a finite set of rational arguments that satisfy the definition of the function. This set should be large to cover enough rational numbers that would appear dense in the diagram, but would not be too close as to merge into a single block. Figure 8.13 gives an idea of the shape of this unusual function.

- Bergfeldt function. This is a little known mapping, discovered by Aksel Bergfeldt [Bergfeldt 18], that maps a real number to a real number, and is everywhere discontinuous. Given a real argument x, we look at its binary value and denote its bits by b_i. The Bergfeldt function is defined by

$$f(x) = \begin{cases} \sum_{k=1}^{\infty} \frac{(-1)^{b_k}}{k}, & \text{if the sum converges,} \\ 0, & \text{otherwise.} \end{cases} \tag{8.1}$$

The sum of Equation (8.1) is a finite part (a tail) of the harmonic series, Section 5.10.1. This series diverges, but reference [Coutteau 17] proves that any tail of it can be made to converge to any real number by changing the signs of some of its terms.

The Bergfeldt function is everywhere discontinuous because it maps every open interval of the real numbers onto the entire real number line (but I couldn't find a proof of this claim).




```
George Beck "The Modified Dirichlet Function"
http://demonstrations.wolfram.com/TheModifiedDirichletFunction/
Wolfram Demonstrations Project
Published: March 7 2011
```

Figure 8.13: The Thomae Function.

■ Conway base-13 function. This function is surjective, and it is this feature which makes it everywhere discontinuous.

A surjective function $y = f(x)$ maps some domain element x to every codomain element y. No y is left behind, but an element y may be the map of several x elements.

As its name implies, this base-13 function is based on tridecimal (base-13) numbers. The representation of such a number requires 13 digits, and the standard convention is to choose the ten decimal digits and the letters A, B, and C. Thus, for example, $601437263_{10} \to 977ACB4B_{13}$ and $6214.38263_{10} \to 2aa0.4c883ac0a1_{13}$.

However, in order to define his uncommon and strange function, Conway substituted the arithmetic symbols "`+`," "`-`," and "`.`" for A, B, and C, respectively. Now, the number 3629265 has the base-13 representation `9+0--7`, but the number 54349589 has a base-13 representation of `-34.128`, which resembles a normal decimal number. The base-13 function is now defined by the following rules:

1. If from some point onwards, the tridecimal expansion of x is of the form $+x_1x_2\ldots x_n.y_1y_2\ldots$, then this number is the value of $f(x)$.

2. If the tridecimal expansion of x ends with $-x_1x_2\ldots x_n.y_1y_2\ldots$, then this negative number is the value of $f(x)$ and everything preceding the minus sign is ignored.

3. In any other case, $f(x) = 0$.

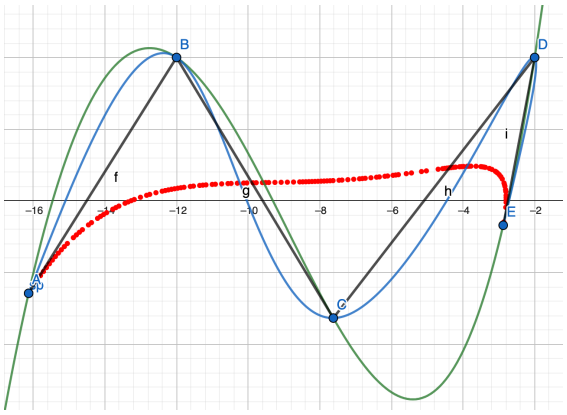
The main reference to this unusual function is [Base 13 function 24]. It shows that this function is everywhere discontinuous, because it is unbounded on every interval around every point. In other words, when applied to all the real numbers in any real interval (a, b), this function returns every real number. As a result it is discontinuous everywhere. Another interesting property is that given an interval on the real number

line, the function maps almost all the real numbers in the interval to zero.

8.4 Interpolation

Typical dictionary definitions of the term interpolation (in mathematics) are "The insertion of an intermediate value or term into a series by estimating or calculating it from surrounding known values," and "A method of constructing (finding) new data points based on the range of a discrete set of known data points." This section concentrates on the latter definition and describes several important methods of interpolation commonly used to construct curves that interpolate discrete sets of points and can be smooth, pleasing to the eye, and also editable (can be reshaped and controlled by the user interactively in real time).

Figure 8.14, made by GeoGebra, illustrates how five points, labeled A through E, are interpolated by (1) straight segments (black), (2) a polynomial (Lagrange, green), (3) a spline (cubic, blue), and (4) a Bézier curve (red dashed). It is clear that the straight segments pass through the points but may not be what a designer would have in mind for the final curve. The Lagrange polynomial is smooth and also passes through the points, but it oscillates wildly and escapes to infinity before it enters and after it leaves the interval of points. Most graphics designers may agree that the blue spline curve is a better interpolation, because it is tight. Finally, the Bézier curve starts at point A and ends at point E, but it doesn't pass through the interior points. Nevertheless, this type of curve is popular and has important applications in many graphics software programs and packages, because it interacts with the designer in an intuitive way.




```
Polynomial({A, B, C, D, E})
Spline({A, B, C, D, E}, 3)
A (1-t)^4 + 4B t(1-t)^3 + 6C t^2(1-t)^2 + 4D t^3(1-t) + E t^4
```

Figure 8.14: Four Interpolation Techniques.

See also Page 446 for bilinear interpolation.

8.5 Least Squares Interpolation

When faced with a set of N points, the challenge of interpolation involves finding a function that smoothly connects the points, without unnecessary fluctuations, abrupt turns, or sudden changes, allowing for accurate estimation of values within the given range. Polynomials are frequently used for interpolation as they can achieve any desired shape and are simple and quick to calculate.

If the degree n of the interpolating polynomial is low (if $n < N$), the best that can be done is to employ the least-squares method to compute a low-degree polynomial that would be the closest possible to all the points. If $n = N$, it is easy to compute an N-degree polynomial that passes through the N points, but such a polynomial tends to oscillate, especially for large values of N. In such a case, a spline curve (a piecewise polynomial, Section 8.13) is the best solution. Polynomials of degrees higher than N are irrelevant to the problem of interpolation.

The least-squares technique deals with the problem of finding the curve that is closest to a given set of points. The term "closest" means that the sum of distances of the points from the curve is the minimum possible for any curve. We start with the simplest curve, a straight line. Figure 8.15a shows a set of $N = 8$ points (x_i, y_i) and a straight line $y = ax + b$ that is close to many of them. The line we are looking for is the one that minimizes the distances from itself to all the points, and it is the term "minimize" that is the source of the word "least" in least squares.

The blue vertical arrows e_i from the line to the points are not the actual distances, which are perpendicular to the line and are computationally difficult to derive, but the much simpler vertical distances. Notice that the distances—which are also referred to as errors, offsets, or residuals—can be negative. The problem has been reduced to finding those values of parameters a and b that minimize the sum $\sum_{k=1}^{N} [y_k - (ax_k + b)]$ of the (vertical) distances.

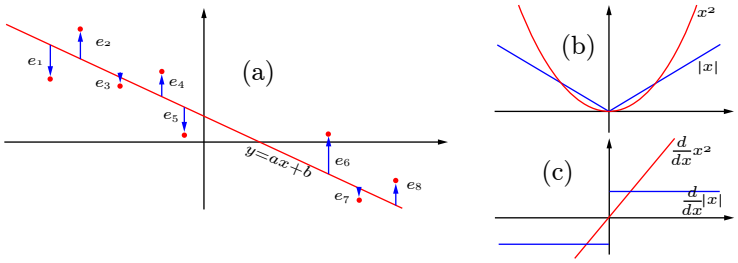


Figure 8.15: A Line Close to Eight Points.

Because the distances e_i are signed, the sum that we need is that of the absolute values of the distances, but in practice we use the sum of the squares of the distances, because of the following:

- The next step is to compute the partial derivatives of the sum with respect to a and b, but the derivative of the absolute value function $y = |x|$ is discontinuous at $x = 0$, because the function has no derivative at this point. This is illustrated in Figure 8.15b,c.

- What we really need is not necessarily the sum of distances, but any quantity that would produce the same values for a and b as those produced by the sum. The sum of squares is such a quantity, which is why the word "square" is part of "least squares." The same values of a and b that minimize the sum of squares also minimize the sum itself, and the derivative of the function $y = x^2$ is linear and continuous (see the red line in part (c) of the figure).

We therefore construct the sum of squares

$$E(a,b) = \sum_{k=1}^{N} [y_k - (ax_k + b)]^2, \tag{8.2}$$

and compute its partial derivatives

$$\begin{aligned}
\frac{\partial}{\partial a}\sum_{k=1}^{N}[y_k-(ax_k+b)]^2 &= \sum_{k=1}^{N}\frac{\partial}{\partial a}[y_k-(ax_k+b)]^2\\
&= \sum_{k=1}^{N} -2x_k(y_k-(ax_k+b)) = -2\sum_{k=1}^{N} x_k e_k.\\
\frac{\partial}{\partial b}\sum_{k=1}^{N}[y_n-(ax_k+b)]^2 &= \sum_{k=1}^{N}\frac{\partial}{\partial b}[y_n-(ax_k+b)]^2\\
&= \sum_{k=1}^{N} -2(y_k-(ax_k+b)) = -2\sum_{k=1}^{N} e_k.
\end{aligned}$$

Equating these derivatives to zero and simplifying, yields the system of two equations in the two unknowns a and b, whose matrix form is

$$\begin{bmatrix} \sum_{k=1}^{N} x_k & N \\ \sum_{k=1}^{N} x_k^2 & \sum_{k=1}^{N} x_k \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{N} y_k \\ \sum_{k=1}^{N} x_k y_k \end{bmatrix}.$$

These equations can also be converted to the more readable form

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \\ x_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} x_1 & x_2 & \dots & x_N \\ 1 & 1 & \dots & 1 \end{bmatrix}^T \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^T, \tag{8.3}$$

where the superscript T stands for matrix transpose.

A 4-point example

Given the four points $(0,4)$, $(1,-1)$, $(2,1)$, and $(3,2)$, we apply the least-squares method to them and compute the straight segment $ax+b$ that is closest to all of them.

The straight segment requires the values of two unknowns a and b, but we have four known quantities that need to be included in the computations, namely the four points. We write the system of four equations with two unknowns

$$L\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 4 \\ -1 \\ 1 \\ 2 \end{bmatrix},$$

and solve it by first multiplying it on the left by the transpose

$$\begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 1 \\ 2 \end{bmatrix},$$

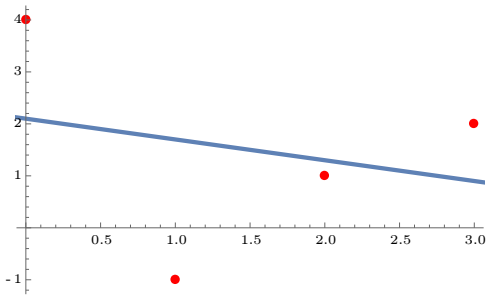
which yields

$$\begin{bmatrix} \frac{1}{5} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{7}{10} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 7 \\ 6 \end{bmatrix}.$$

This is solved by matrix inversion to produce

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{1}{5} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{7}{10} \end{bmatrix}^{-1} \begin{bmatrix} 7 \\ 6 \end{bmatrix} = \begin{bmatrix} -2/5 \\ 21/10 \end{bmatrix}.$$

Figure 8.16 shows the points and the resulting line $y = -\frac{2}{5}x + \frac{21}{10}$.




```
pts={{0,4},{1,-1},{2,1},{3,2}};
L={{0,1},{1,1},{2,1},{3,1}};
LT=Transpose[L].L
Solve[LT.{{a},{b}}==Transpose[L] .{{4},{-1},{1},{2}},{a,b}]
p1=ListPlot[pts,PlotStyle->{PointSize[Large],Red}];
p2=Plot[-(2/5)x+(21/10),{x,-3,5}];
Show[p1,p2]
```

Figure 8.16: Linear Least-Squares Example.

Linear LS with higher-order polynomials

If the given set of points to be interpolated is very different from a straight line, the least-squares technique can still be used to fit a curve close to the points by increasing the degree of the curve. Mathematically, this starts by extending Equation (8.2) from a straight line (a degree-1 polynomial) to a degree-n polynomial. This produces the error function

$$E(a_1, a_2, \ldots, a_n) = \sum_{k=1}^{N} [y_k - (a_1 x_k + a_2 x_k^2 + \cdots + a_n x_k^n)]^2. \tag{8.4}$$

This is still considered linear least squares, because the coefficients a_i of the polynomial appear linearly and not, for example, as a_i^2 or a_i^3.

The n partial derivatives of Equation (8.4) with respect to the coefficients a_i are then calculated and equated to zero, producing a final result that is a direct extension of Equation (8.3)

$$\begin{bmatrix} x_1^{n-1} & x_2^{n-1} & \ldots & x_N^{n-1} \\ \vdots & \vdots & & \\ x_1^2 & x_2^2 & \ldots & x_N^2 \\ x_1 & x_2 & \ldots & x_N \\ 1 & 1 & \ldots & 1 \end{bmatrix}^T \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_n \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \ldots & y_N \end{bmatrix}^T, \tag{8.5}$$

The degree-n polynomial is then computed and plotted with the a_i coefficients that are the solutions of Equation (8.5).

A degree-2 four-point example

Given the four points $(0, 4)$, $(1, -1)$, $(2, 1)$, and $(3, 2)$, we apply the least-squares method to determine the degree-2 polynomial (a parabola) $p(x) = ax^2 + bx + c$ that best fits them. There are now three unknowns a, b, and c, and we first construct the system of four equations

$$L \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 4 \\ -1 \\ 1 \\ 2 \end{bmatrix}.$$

Notice that the first column of matrix L is the square of its second column. This is because the leftmost parameter, a, is associated with x^2, the second parameter, b, is multiplied by x, and the third parameter, c is the coefficient of 1. This system of equations is solved by first multiplying it on the left by the transpose

$$\begin{bmatrix} 0 & 1 & 4 & 9 \\ 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 & 1 & 4 & 9 \\ 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 1 \\ 2 \end{bmatrix},$$

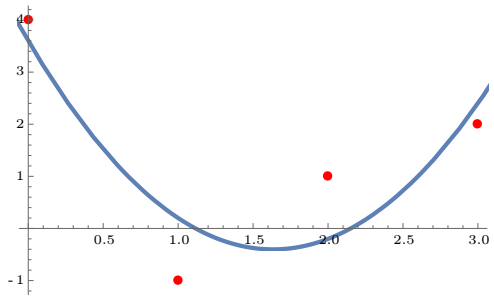
which yields

$$\begin{bmatrix} 98 & 36 & 14 \\ 36 & 14 & 6 \\ 14 & 6 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 21 \\ 7 \\ 6 \end{bmatrix}.$$

This is solved by matrix inversion to produce

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 98 & 36 & 14 \\ 36 & 14 & 6 \\ 14 & 6 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 21 \\ 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 3/2 \\ -49/10 \\ 18/5 \end{bmatrix}.$$

Figure 8.17 shows the points and the resulting parabola $y = \frac{3}{2}x^2 - \frac{49}{10}x + \frac{18}{5}$.




```
pts={{0,4},{1,-1},{2,1},{3,2}};
L={{0,0,1},{1,1,1},{4,2,1},{9,3,1}};
LT=Transpose[L].L
Solve[LT.{{a},{b},{c}}==Transpose[L] .{{4},{-1},{1},{2}},{a,b,c}]
p1=ListPlot[pts,PlotStyle->{PointSize[Large],Red}];
p2=Plot[(3/2)x^2-(49/10)x+18/5,{x,-3,5}];
Show[p1,p2]
```

Figure 8.17: A Least-Squares Parabola.

Figure 8.18 shows the two previous examples computed and plotted quickly in Geogebra. The green curve is a degree-3 polynomial. Such a polynomial has four coefficients, which is why its fit is perfect. Instead of getting close to the points, the resulting system of four equations with four unknowns has a unique solution, the degree-3 polynomial which passes through the points.

The least squares method described so far is linear. This is because the coefficients a_i of the interpolating polynomial appear in Equation (8.4) linearly and not, for example, as a_i^2 or a_i^3. There is also a nonlinear version of least squares, and the following is a quotation from reference [nonlinearLS 24].

> Non-linear least squares is the form of least squares analysis used to fit a set of m observations with a model that is non-linear in n unknown parameters ($m \geq n$). It is used in some forms of nonlinear regression. The basis of the method is to approximate the model by a linear one and to refine the parameters by successive iterations. There are many similarities to linear least squares, but also some significant differences.

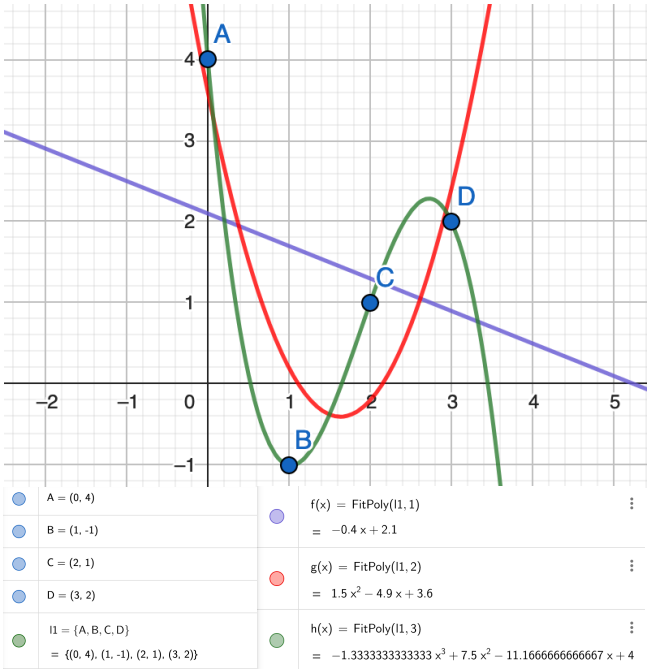


Figure 8.18: Parametric Fit in GeoGebra.

8.6 Perlin Noise

Perlin noise is an important type of a procedural texture. It is used in practice in computer graphics to enhance realism in still images and videos.

Noise as Texture

Just like various techniques, formulas, and approaches in software development, Perlin noise relies on utilizing random numbers. Nevertheless, because software is deterministic and operates based on rules, the supposedly random number sequences it produces are not truly random. They seem haphazard and meet the criteria for randomness when tested statistically, which is why they are called pseudo-random (refer also to the conversation on Page 356).

Definitions
Random noise refers to any kind of noise that is unpredictable and has no pattern. It is essentially a random signal that can have any frequency or amplitude. White noise, on the other hand, is a type of random noise that contains the same energy at any interval of frequencies. It has a flat spectrum. When played as audio, it sounds like static or hissing.

The term “noise” has many meanings and is used in many contexts. Section 10.4 mentions noise in connection with error-control codes, but Perlin noise is an example

of noise as texture. Perlin noise is an algorithm to create texture in digital images. Wikipedia defines image texture as "the small-scale structure perceived on an image, based on the spatial arrangement of color or intensities."

Texture can be included in an image by scanning a natural or artificial texture and adding it to the image. However, researchers have developed algorithms to generate various types of texture by software, and this kind of texture is referred to as procedural or solid texture. Including natural texture in an image increases the image file size and may also involve a mapping problem. A typical mapping problem is how to wrap a two-dimensional texture around a three-dimensional complex digital object, so that the object still looks real. Using procedural texture eliminates the mapping problem, saves on image file size, but requires extra processing.

Noise added to an image as texture acts like spice added to bland food. It contributes a certain amount of randomness and irregularity to the image that, if handled properly, adds interest to the image and makes it come alive. Traditional painters who used oil paints were also aware of this feature and tried to make irregular brush strokes that produced random paint textures on the canvas.

Mathematically, noise as texture is a mapping whose input is the coordinates of an n-dimensional point and whose output is a real number, the noise value at the point (color, shade of gray, or height). A one-dimensional noise is a sequence of real numbers a_i produced in points t_i in time, while a two-dimensional noise is an array of real numbers, each corresponding to a pair (x, y) of coordinates. Such noise can be viewed as a random pattern of color or grayscale dots on a plane. It can also be interpreted as the heights of points above a surface and thus to describe a terrain.

A three-dimensional noise is an array of real numbers, each corresponding to a triplet (x, y, z) of coordinates. It can correspond, for example, to the colors of points in a solid three-dimensional object or it can be used to slightly perturb the position of each object point in space, thereby introducing roughness to the object.

Perlin noise is the brainchild of Ken Perlin, who started developing it in the early 1980's, published it in 1985, and used it in the Disney movie *Tron*. Perlin was unsatisfied with the noise-generation methods available to him, because they produce decorrelated noise.

Formally, Perlin noise is a gradient noise. It starts with a set of pseudo-random gradients placed at regularly-spaced points in n-dimensional space, and then a smooth function is computed by interpolating between the points.

Figure 8.19 illustrates this kind of noise both in one- and two dimensions. It is easy to see how adjacent points have very different heights (in 1D) or shades of gray (in 2D). A point's location or color can be very different from its neighbors' and is independent of them. Perlin decided to develop correlated noise, a type of noise where the gray intensity or color of a point is never very different from those of its four (up, down, left, right) or even eight nearest neighbors.

One-Dimensional Correlated Noise

We Start with an intuitive explanation of one-dimensional correlated noise. The middle part of Figure 8.19 (three octaves of correlated random 1D noise) was prepared by reference [Perlin 1D 24]. It shows 12 initial random points that are approximated by a smooth curve, but are decorrelated. The height of a point may be very different

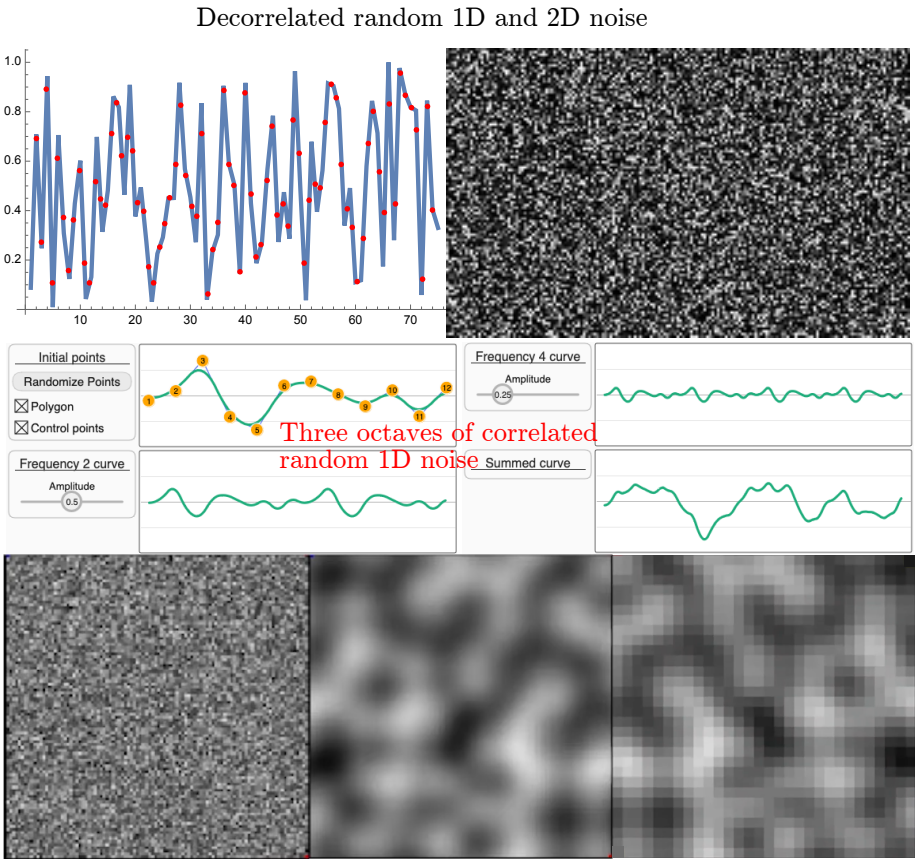


Figure 8.19: Correlated and Decorrelated 1D and 2D Noise.

from those of its neighbors. This pattern is then transformed by doubling its frequency and halving its amplitude, and is placed underneath the original graph. It becomes a new octave. This octave generation process is repeated again to create a copy with four times the frequency of the original and one-fourth its amplitude. Finally, the original and its two octaves are added up. The result is a curve that looks random and smooth. It no longer approximates the 12 original points, but each real point on this curve is at a similar height to its near neighbors.

This close relation between nearby points is the essence of correlated noise. The number of octaves and their shrink ratios can be varied, but the process, which is termed "sum $1/f$ noise," is the same.

Note. Octaves can be generated and added until the size of the next one becomes too small to see. This happens when its noise fluctuates faster than about once every two pixels.

Two-Dimensional Perlin Noise

The principle of creating two-dimensional correlated noise has been described in

Perlin's original 1985 publication, reference [Perlin 85], and has been improved, extended, described, explained, and published in numerous scientific papers, articles, talks, and videos in various venues.

The algorithm consists of two main parts. In the first part, described here in detail, noise values are computed for individual pixels. The second part is optional and is similar to the octaves described above. The array of noise values is rescaled, copied, and added to itself in steps, where its frequency is doubled and its amplitude is halved. This results in highly correlated noise which can also be described as fractal, because each part of the noise is similar to the entire region of noise. When a small part of the noise is copied and is pasted over another area, a viewer cannot perceive a change in the overall noise pattern, which is normally not true for decorrelated noise.

It is possible to obtain beautiful, unexpected results by playing with octaves. An octave can be reflected, scaled, perturbed slightly and randomly, or converted to absolute values. An "artificial" low-frequency sine wave octave can be prepared and added to the other octaves. Attractive animations can easily be produced by combining two noises linearly, as in Equation (8.6).

Using the absolute values of octaves produces exciting results, especially if combined with red-orange colors. The following is a quotation from [Perlin 99]:

> The application of the absolute value causes a bounce or crease in the noise function in all the places where its value crosses zero. When these modified noise functions are then summed over many scales, the result is visual cusps—discontinuities in gradient—at all scales and in all directions. The visual appearance is consistent with a licking flame-like effect, if it's colored properly. In 1984 I started calling this formulation turbulence, since it gives an appearance of turbulent flow.

The second part of the algorithm starts with a two-dimensional rectangular region of points and it computes a noise value—height, color, or grayscale—for each point in the region. The main steps of the method are (1) Generating cells and gradient vectors. (2) Computing four dot products for each pixel of a cell. (3) Interpolating the four dot products of a point to produce the noise value at the point.

The region to be filled with Perlin (i.e., correlated) noise is first partitioned into $m \times n$ square cells of pixels, and a (pseudo) random gradient vector is placed at the four corners of each cell. Each vector is shared by four cells, except the ones on the sides and at the corners of the region. The total number of gradient vectors is $(m+1) \times (n+1)$ and they remain fixed during the computations.

The term "gradient" means a gradual change from one value to another. In principle, the gradient vectors are random, but later research showed how they can be chosen cleverly. Also, the gradient vectors can be of any length, but should all have the same length, because this makes it possible to figure out in advance the range of the final noise values.

Note. This description is for the two-dimensional case. In three dimensions, each cell would be a cube with eight three-dimensional gradient vectors at its corners, and each vector would be shared by eight cubes. In general, there would be a gradient vector at each of the 2^n corners of an n-dimensional hypercube.

Each cell is square and consists of rows and columns of pixels, where a noise value has to be calculated for each pixel. The algorithm loops over the pixels of a cell and for

each pixel it computes a distance (or offset) vector from each corner of the cell to the pixel. The units of the distance vectors are simply the coordinates of the pixel from the top-left of the cell.

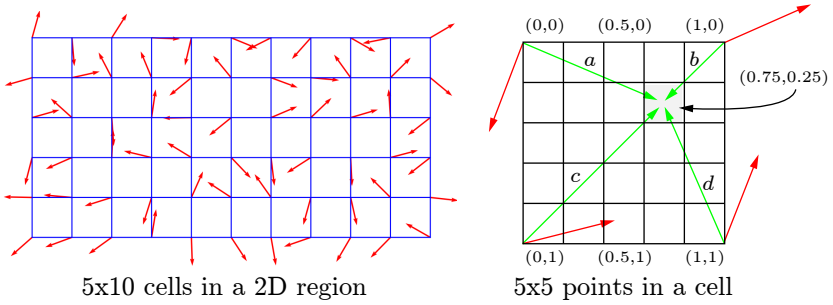


Figure 8.20: Gradient and distance Vectors.

Figure 8.20 shows a region partitioned into 5×10 cells with 6×11 gradient vectors (in red). One cell is shown, consisting of 5×5 pixels, with four distance vectors (in green) pointing to the pixel at $(0.75, 0.25)$ in the cell. For the purpose of our calculations we assume that the cell is a square, one unit on a side, with an origin at its top-left corner. x values vary from 0 on the left to 1 on the right, and y values vary from 0 at the top to 1 on the bottom. The four distance vectors are therefore $a = (0.75, 0.25)$, $b = (-0.25, 0.25)$, $c = (0.75, -0.75)$, and $d = (-0.25, -0.75)$.

The dot products are next. See Page 472 and Equation (8.22) for more on the dot product, and keep in mind that dot products are real numbers and can be negative.

Once all eight vectors are known, the four dot products are easy to compute and they become the four gradient values for the pixel. The dot product of two vectors is the product of their lengths times the cosine of the angle between them. The distance vectors always point inside the cell and are at most one unit long, but the gradient vectors can be of any length, and it is obvious that their lengths affect the dot products and through them, the final noise value.

It is therefore preferable to have gradient vectors of the same length. Denoting the length of the gradient vectors by l and those of the distance vectors by d, the dot product equals $l \cdot d \cdot \cos\theta$. The lengths d of the distance vectors vary from zero to the diagonal of the unit cell, so they are in the interval $[0, \sqrt{2}]$ (for n-dimensional noise, the diagonal is $\sqrt{n}$). The cosine is always in the interval $[-1, 1]$, which implies that the dot products would be in the interval $[-\sqrt{2}l, \sqrt{2}l]$, and this would also be the interval of the noise values, unless they are modified by a smoothing function as discussed later. Also, adding octaves increases the noise values.

We are now ready to see how the dot products of gradient vectors and distance vectors create smooth blobs with white, gray, and black areas. We assume that our gradient vectors are unit vectors and therefore contribute only their directions to the dot products. The value of a dot product in this case is affected only by (1) the length of the distance vector, and (2) the angle between the vectors. The greater the length and the smaller the angle, the larger the dot product.

Figure 8.21 shows a region of 2×5 cells and we assume that each cell consists of many pixels. The row of blue pixels is spread in the direction of a red gradient vector, so for those pixels the angle is zero and the dot products are simply the lengths of the distance vectors. As we move away along the blue pixels from the cell's corner, those lengths start at zero and get longer as the pixels get deeper into the cell. We now convert the dot product values to grayscale with 0 becoming white and 1 becoming black. This justifies the shading in the figure, which varies from dark to white along the blue pixels. (This also justifies the name "gradient vectors." The direction of such a vector corresponds to the color gradient along a line of pixels.)

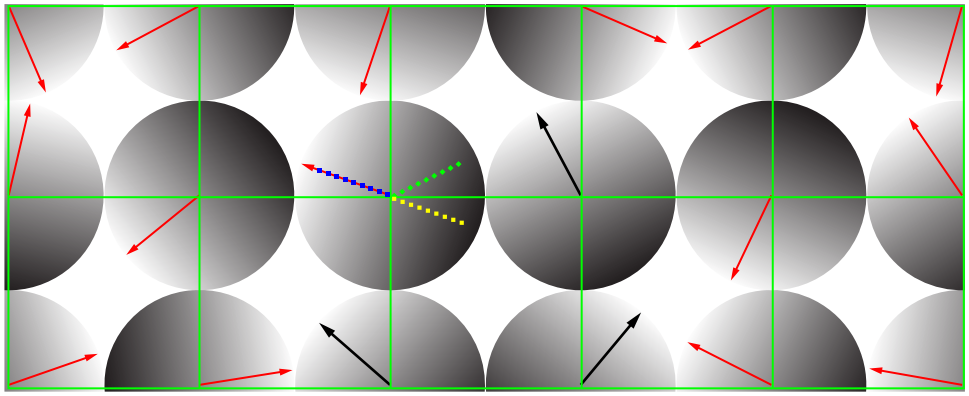


Figure 8.21: Dot Products Converted into Grayscales.

The situation with the yellow pixels is similar, the only difference being the 180° angle between their distance vectors and the red gradient vector, which implies a zero cosine. They should therefore be all black, but they are not because the final color of those pixels is also affected by the other three (thick, black) gradient vectors of the cell. The angle between the distance vectors of the green pixels and the red gradient vector is about 150°, which is why they should all be dark, but their final color is also affected by the same three thick, black gradient vectors of the same cell. The final noise value (color, shade of gray, or height) of a pixel is a mixture or interpolation of the noise values produced by four dot products.

One more practical point to consider is the lengths of the distance vectors. They can vary from zero to the diagonal of the cell, which is why an actual implementation of Perlin noise should scale those lengths to the interval $[0, 1]$.

This is how the dot products start the construction of the random grayscales of the pixels. The construction is completed by the bilinear interpolation described next, but before we turn to that topic let's consider the important feature of Perlin noise namely, that it is correlated. It is now clear that the noise is correlated, the grayscales of adjacent pixels never vary much, because the dot products depend on both the distance vectors and the angles, and these two quantities vary little between adjacent pixels.

The last major step is to interpolate the four dot products for each point. We start with a detailed description of bilinear interpolation.

Bilinear Interpolation

We are familiar with the concept of an average. The average of the numbers a and b is $(a+b)/2$, half their sum. This can be extended to more than two numbers by considering an average a special case of the general weighted sum $\alpha a + \beta b$. An average is the case $(0.5a + 0.5b)$, where the two numbers are given equal weights of 0.5 each. In the general weighted sum, the two numbers are given different weights, but it doesn't take long to realize that such a weighted sum makes sense only if the weights add up to 1, in which case the weighted sum takes the form

$$(1-t)a + tb. \tag{8.6}$$

When t is varied in the interval $[0,1]$, the sum varies linearly from a (for $t=0$) to b (for $t=1$). The term "linearly" is used because when Equation (8.6) is plotted as a function of t, the result is a straight line segment and the equation is the basic parametric equation of a straight segment.

This connection between average and straight segments is now extended and is used to develop the technique of bilinear interpolation. The particular form of bilinear interpolation that we need is illustrated in Figure 8.22.

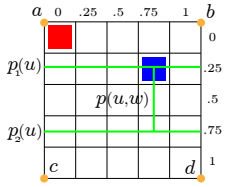


Figure 8.22: Bilinear Interpolation of Pixels in a Cell.

Four numbers a, b, c, and d need be averaged. They are assigned to the four corners of a 5×5 cell of pixels. We want to compute a weighted average of the four numbers for each of the 25 pixels, such that the closer a pixel is to a corner, the more weight the corner gets in the average. The weighted average for the pixel nearest corner a would be a itself, while the weighted average for the central pixel would assign a weight of 0.25 to each corner. This is the essence of bilinear interpolation, a process that is performed in the following three steps:

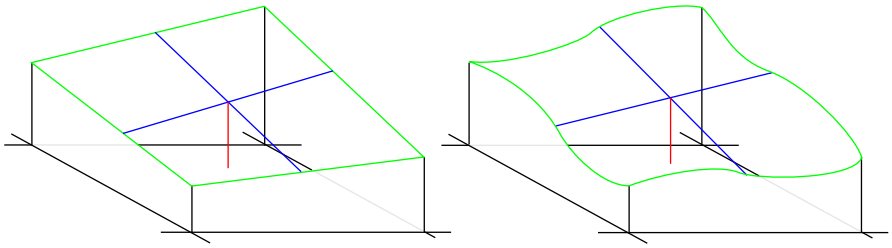
- Construct the parametric equation $p_1(u) = (1-u)a + ub$ of the straight segment between a and b.
- Construct the parametric equation $p_2(u) = (1-u)c + ud$ of the straight segment between c and d.
- Construct the parametric equation $p(u,w) = (1-w)p_1(u) + w\,p_2(u)$ of the straight segment between an arbitrary point on $p_1(u)$ and the point on $p_1(u)$ that corresponds to the same u value. The result is

$$p(u,w) = (1-w)p_1(u) + wp_2(u) = (1-w)(a - ua + ub) + w(c - uc + ud)$$

$$
\begin{aligned}
&= a - wa - ua + uwa + ub - uwb + wc - uwc + uwd \\
&= a + u(b-a) + w(c-a) + uw(a-b-c+d) \\
&= a(1-u)(1-w) + b\,u(1-w) + c\,w(1-u) + d\,uw. \qquad (8.7)
\end{aligned}
$$

Equation (8.7) can now be used to compute the bilinear interpolation of the four corner values for each of the 25 pixels in the cell, by assigning u and w the coordinates of a pixel relative to corner a. Thus, the red pixel in Figure 8.22 has coordinates $(0, 0)$, the blue pixel has coordinates $(0.75, 0.25)$ and the center pixel's coordinates are $(0.5, 0.5)$. The three parametric segments are shown in green.

Figure 8.23a illustrates a typical bilinear interpolation of four dot product values. It shows a square with four vertical (black) lines at its corners and four straight (green) segments connecting them. Two blue lines connect the centers of the green segments, and the red vertical line becomes the bilinear interpolation of four vertical (black) lines. The red line represents the height of the center of the green square above the black square.



(a) (b)

Figure 8.23: Bilinear Interpolation of Four Dot Products.

Early experiments with the Perlin noise algorithm produced unsatisfactory noise that was not smooth (i.e., not correlated) enough. Much improved (correlated) noise was produced when the basic bilinear interpolation was enhanced by a smoothing function. The first tests used the "ease" smoothing function $3t^2-2t^3$. This function is the Hermite blending (or weight) function F_2, listed in Equation (8.31) and displayed in Figure 8.49. It maps 0 to 0, 1 to 1, and 0.5 to 0.5, and is shaped like the letter `S`, which is why it magnifies values larger than 0.5, while reducing smaller values. Figure 8.23b shows a typical set of four `S`-shaped curves (in green) arranged over a square box for better bilinear interpolation.

Even better results are achieved with the "fade" smoothing function

$$6t^5 - 15t^4 + 10t^3 = t \times t \times t \times (t \times (t \times 6 - 15) + 10).$$

In addition to mapping $0 \to 0$, $1 \to 1$ and $0.5 \to 0.5$, this `S`-shaped function is smoother because both its first and second derivative are zero at $t = 0$ and $t = 1$.

◇ **Exercise 8.6:** Derive the "ease" smoothing function $3t^2 - 2t^3$ as a linear interpolation of two parabolas.

Figure 8.24 shows the two smoothing functions together with typical direct and smoothed bilinear interpolations.

Other smoothing functions are possible. For example, $\sin[\pi(t - 0.5)] + 0.5$ also maps 0, 0.5, and 1 to themselves and produces a nice S-shaped curve.

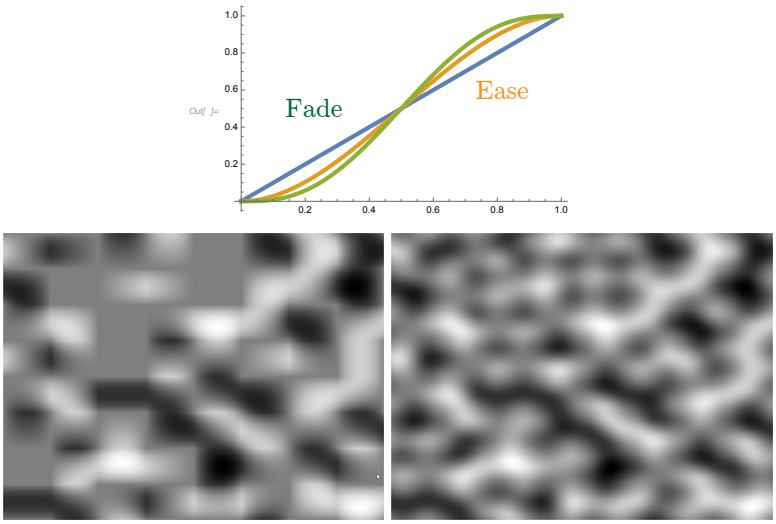


Figure 8.24: Bilinear Interpolation: Direct (Left) and Smooth (Right).

Choosing the Gradient Vectors

It turns out that the gradient vectors don't have to be random and don't have to be unique. They are used to generate noise, and noise is inherently random. Two techniques were used by Perlin for choosing these vectors.

- The original method employed by Perlin was to choose two random numbers, consider them the coordinates of a two-dimensional vector, and normalize them to end up with a unit vector. This is readily extended to n-dimensional noise. It was later realized, however, that this method produces more vectors pointing along or close to the diagonals and fewer vectors pointing along or close to the cell edges.

The reason for this unexpected behavior is similar to what happens in dice rolling. Throwing a single die produces one of the integers 1 through 6 with equal probabilities, but rolling two dice produces the 11 results $1 + 1 = 2$ through $6 + 6 = 12$ with unequal probabilities. The value 7 is produced when rolling $(1, 6)$, $(2, 5)$, and $(3, 4)$, while 2 is produced only for $(1, 1)$. Here are two ways to solve this problem:

1. Pick two random numbers that are normally distributed (i.e., distributed on a bell curve) and combine them into a vector. For n-dimensional noise, choose n such random numbers to become the coordinates of a vector.

2. (This only works for two dimensions.) Instead of choosing two random numbers for a vector (which results in a cubic distribution of the vectors), random angles $\theta \in [0, 2\pi]$ should be chosen, and each used to generate one random gradient vector $(\cos\theta, \sin\theta)$. This results in a spherical distribution of the vectors.

- An improved version, also developed by Perlin. Construct a fixed set of vectors. The first time the noise algorithm requires a gradient vector at a corner, pick one up from the set at random, use it, and remember it. When coming back to the same corner again, choose the same gradient vector.

While constructing such a set, the vectors can be created normalized, but an alternative is to generate vectors with simple components whose size would not be 1, but whose simple components would speed up the calculations. For two-dimensional noise, the set of vectors whose components are 1, -1, and 0 are especially useful and simplify the computations of the dot products. This set consists of the eight vectors $(-1,-1)$, $(-1,0)$, $(-1,1)$, $(0,-1)$, $(0,1)$, $(1,-1)$, $(1,0)$, and $(1,1)$. Perlin proposes a similar set for the three-dimensional case.

The dot product is at its maximum when its two inputs are parallel. Imagine a gradient vector that happens to be almost parallel to an edge of its cell. The dot products of such a vector with the pixels along that edge would be large, because of the small angles, and the opposite is true for the dot products of this gradient vector with the pixels on the same edge but in the adjacent cell (where the angles would be close to 180°). The result would be several small dot products followed by several large ones. This implies several small noise values followed by several large values.

Now imagine that several gradient vectors in neighboring cells also happen to be aligned in the same direction. Such a situation may cause many large blobs of bright noise followed by dark noise to concentrate in a small area, producing a regular effect. Therefore , instead of setting the gradient vectors to point in random directions, any implementation should make sure that no adjacent gradient vectors point parallel to an edge. This requires much experimentation, and Ken Perlin has come up with a permutation table that can be used by software to distribute the gradient vectors in a configuration that eliminates this problem.

Figure 8.25 shows why this problem exists only for gradient vectors that point close to an edge. A gradient vector that points to the center of its cell would result in a large dot product only for the center pixels in the cell, but this group of pixels is not adjacent to the center pixels in adjacent cells.

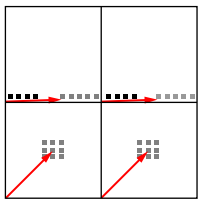


Figure 8.25: Repetition in Gradient Vectors.

Value Noise and Gradient Noise

The Perlin noise described above is an improved version [Perlin 01] of Perlin's original 1985 algorithm, and is referred to as gradient noise. A simpler version of this method is known as value noise. This is a faster algorithm whose output tends to look blocky and is often deemed unsatisfactory. Figure 8.26 compares typical value noise and gradient noise. Value noise starts by assigning random values instead of random gradient vectors to all the corner cells. No dot products are computed, and bilinear interpolation (or a smooth version of it) is applied to the four random values at the corners of a cell instead of to the results of four dot products.

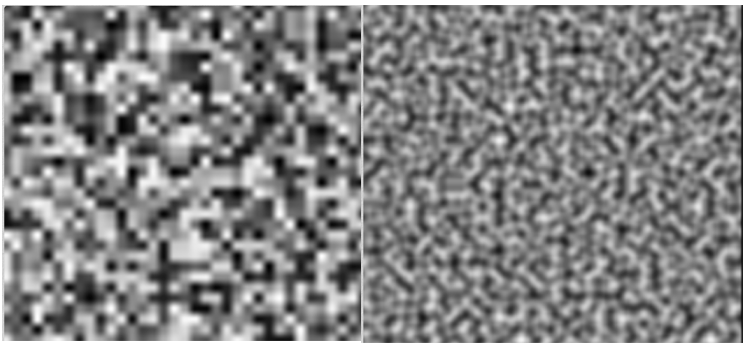


Figure 8.26: Value Noise (Left) and Gradient Noise (Right).

8.7 Three-Dimensional Rotations

The following sections ares devoted to rotations in three-dimensions. Because of the confusing nature of rotations in three dimensions, the discussion starts with vector operations, methods to implement rotations, and algorithms to speed up the computations. Section 3.6, on two-dimensional rotations is a good introduction that should be covered before reading ahead. Those who want to go beyond what is discussed here are invited to peruse Section 2.9 for a description of quaternions and how they relate to rotations.

A rotation is a transformation (of a point, an object, or of space) that preserves at least one point. A rotation has a magnitude, measured in radians or degrees, and a direction such as clockwise. In contrast, a translation moves every point of the translated object, and reflections leave fixed either a line (in two dimensions) or a plane (in three dimensions).

In two dimensions, rotations are about a fixed point, the center of rotation. In three dimensions, rotations are about a fixed axis. Mathematically, a rotation is a mapping from real points to real points. All the possible two-dimensional rotations about the same center form a group (a rotation group), and similarly for rotations in three dimensions.

This section starts with a survey of several vector operations that are needed later. It continues with discussions of the direction of three-dimensional rotations, Basic rotations, general rotations, coordinate-free rotations, the CORDIC algorithm, and quaternions. The large amount of material covered here is an indication of how complex and unintuitive this topic is.

8.8 Vector Operations

We start with a review of several vector-related terms.

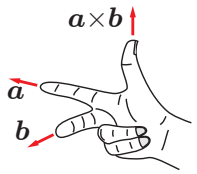
• The norm of a vector $\mathbf{v} = (v_x, v_y, v_z)$ is its magnitude. It is denoted by $\|\,\mathbf{v}\,\|$ and its value is $\sqrt{v_x^2 + v_y^2 + v_z^2}$.

• The dot product (or scalar product) of two vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ is a number. Algebraically, the dot product, denoted by $\mathbf{a}\cdot\mathbf{b}$, is the sum $a_1b_1 + a_1b_1 + \cdots + a_nb_n$. Geometrically, it is the product

$$\|\,\mathbf{a}\,\|\,\|\,\mathbf{b}\,\|\,\cos(\theta).$$

This definition implies that the dot product of parallel vectors is the product of their magnitudes, and the dot product of two perpendicular vectors is zero.

• The cross-product (or vector product) of two three-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\mathbf{a}\times\mathbf{b}$ and is a vector $\mathbf{c}$ that is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$. The direction of $\mathbf{c}$ is given by the right-hand rule, which is illustrated in the figure. The magnitude (norm) of $\mathbf{c}$ is



$$\|\,\mathbf{a}\,\|\,\|\,\mathbf{b}\,\|\,\sin(\theta),$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{b}$. This magnitude also equals the area of the parallelogram spanned by the two vectors. The cross-product is defined only for three-dimensional vectors.

The cross product can also be represented as the vector

$$\|\,\mathbf{a}\,\|\,\|\,\mathbf{b}\,\|\,\sin(\theta)\,\mathbf{n} = (a_2b_3 - a_3b_2, -a_1b_3 + a_3b_1, a_1b_2 - a_2b_1), \tag{8.8}$$

where $\mathbf{n}$ is a unit vector perpendicular to the plane containing $\mathbf{a}$ and $\mathbf{b}$. Its direction is such that the ordered set $(\mathbf{a}, \mathbf{b}, \mathbf{n})$ is positively oriented (i.e., a screw driven from $\mathbf{a}$ to $\mathbf{b}$ would move in the positive direction of $\mathbf{n}$). It is easy to show that $\mathbf{a}\times\mathbf{b}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$.

◇ **Exercise 8.7:** Show it!

Perhaps the best proof is to construct the cross product from first principles. Given the two vectors $\mathbf{P}$ and $\mathbf{Q}$, we are looking for a vector $\mathbf{R}$ perpendicular to both $\mathbf{P}$ and $\mathbf{Q}$. This requirement does not fully define $\mathbf{R}$, since both $\mathbf{R}$ and $-\mathbf{R}$ satisfy it and since it says nothing about the magnitude of $\mathbf{R}$. We therefore extend our definition of the cross product by requiring that the triplet $(\mathbf{P}, \mathbf{Q}, \mathbf{R})$ be a right-handed triad of vectors and also that the magnitude of $\mathbf{R}$ be the product $\|\,\mathbf{P}\,\|\,\|\,\mathbf{Q}\,\|\,\sin\theta$, where θ is the angle between $\mathbf{P}$ and $\mathbf{Q}$.

The derivation exploits the orthogonality of the three coordinate axes $\mathbf{i} = (\mathbf{1}, \mathbf{0}, \mathbf{0})$, $\mathbf{j} = (\mathbf{0}, \mathbf{1}, \mathbf{0})$, and $\mathbf{k} = (\mathbf{0}, \mathbf{0}, \mathbf{1})$ and also uses our definition. The definition implies that $\mathbf{i}\times\mathbf{i} = \mathbf{0}$ because the angle between $\mathbf{i}$ and itself is zero, and the same holds for $\mathbf{j}$ and $\mathbf{k}$. It also implies that the cross product of any two of the three basis vectors is a unit vector because the basis vectors are unit vectors and because $\sin 90^\circ = 1$. Once we realize that

the triplet $(\mathbf{i},\mathbf{j},\mathbf{k})$ is a right-handed triad, we can deduce the following: $\mathbf{i}\times\mathbf{j}=\mathbf{k}$, $\mathbf{j}\times\mathbf{k}=\mathbf{i}$, $\mathbf{k}\times\mathbf{i}=\mathbf{j}$, $\mathbf{j}\times\mathbf{i}=-\mathbf{k}$, $\mathbf{k}\times\mathbf{j}=-\mathbf{i}$, and $\mathbf{i}\times\mathbf{k}=-\mathbf{j}$.

Armed with this information, we can easily derive the cross product $\mathbf{R}$:

$$\begin{aligned}\mathbf{R}=\mathbf{P}\times\mathbf{Q}&=(P_1\mathbf{i}+P_2\mathbf{j}+P_3\mathbf{k})\times(Q_1\mathbf{i}+Q_2\mathbf{j}+Q_3\mathbf{k})\\&=(P_1\mathbf{i}+P_2\mathbf{j}+P_3\mathbf{k})\times Q_1\mathbf{i}+(P_1\mathbf{i}+P_2\mathbf{j}+P_3\mathbf{k})\times Q_2\mathbf{j}+(P_1\mathbf{i}+P_2\mathbf{j}+P_3\mathbf{k})\times Q_3\mathbf{k}\\&=(P_2Q_3-P_3Q_2)\mathbf{i}+(-P_1Q_3+P_3Q_1)\mathbf{j}+(P_1Q_2-P_2Q_1)\mathbf{k}\\&=(P_2Q_3-P_3Q_2,-P_1Q_3+P_3Q_1,P_1Q_2-P_2Q_1).\end{aligned}$$

The magnitude of $\mathbf{R}$ can be calculated explicitly

$$\begin{aligned}\parallel\mathbf{R}\parallel^2&=(P_2Q_3-P_3Q_2)^2+(-P_1Q_3+P_3Q_1)^2+(P_1Q_2-P_2Q_1)\\&=(P_1^2+P_2^2+P_3^2)(Q_1^2+Q_2^2+Q_3^2)-(P_1Q_1+P_2Q_2+P_3Q_3)^2\\&=\parallel\mathbf{P}\parallel^2\parallel\mathbf{Q}\parallel^2-(\mathbf{P}\cdot\mathbf{Q})^2=\parallel\mathbf{P}\parallel^2\parallel\mathbf{Q}\parallel^2-(\parallel\mathbf{P}\parallel|\mathbf{Q}\parallel\cos\theta)^2\\&=\parallel\mathbf{P}\parallel^2\parallel\mathbf{Q}\parallel^2(1-\cos^2\theta)=\parallel\mathbf{P}\parallel^2\parallel\mathbf{Q}\parallel^2\sin^2\theta.\end{aligned}$$

To illustrate the magnitude, we can draw the parallelogram defined by $\mathbf{P}$ and $\mathbf{Q}$ (with an angle θ between them) and show that vector $\mathbf{Q}\sin\theta$ is perpendicular to $\mathbf{P}$.

The following expressions show how $\mathbf{P}\times\mathbf{Q}$ can be expressed by means of a determinant,

$$\begin{aligned}\mathbf{P}\times\mathbf{Q}&=\begin{vmatrix}\mathbf{i}&\mathbf{j}&\mathbf{k}\\P_1&P_2&P_3\\Q_1&Q_2&Q_3\end{vmatrix}=\mathbf{i}\begin{vmatrix}P_2&P_3\\Q_2&Q_3\end{vmatrix}-\mathbf{j}\begin{vmatrix}P_1&P_3\\Q_1&Q_3\end{vmatrix}+\mathbf{k}\begin{vmatrix}P_1&P_2\\Q_1&Q_2\end{vmatrix}\\&=(P_2Q_3-P_3Q_2,-P_1Q_3+P_3Q_1,P_1Q_2-P_2Q_1),\end{aligned}$$

or, alternatively, by means of a matrix

$$\mathbf{P}\times\mathbf{Q}=(Q_1,Q_2,Q_3)\begin{pmatrix}0&P_3&-P_2\\-P_3&0&P_1\\P_2&-P_1&0\end{pmatrix}. \tag{8.9}$$

◇ **Exercise 8.8:** The cross product $\mathbf{P}\times\mathbf{Q}$ is perpendicular to both $\mathbf{P}$ and $\mathbf{Q}$. In what direction does it point?

The cross product is not commutative and is not associative. It is, however, distributive with respect to addition or subtraction of vectors. Hence, $\mathbf{P}\times(\mathbf{Q}\pm\mathbf{T})=\mathbf{P}\times\mathbf{Q}\pm\mathbf{P}\times\mathbf{T}$.

The magnitude of $\mathbf{P}\times\mathbf{Q}$ equals $\parallel\mathbf{P}\parallel\parallel\mathbf{Q}\parallel\sin\theta$, where θ is the angle between the two vectors. The cross product therefore has a simple geometric interpretation. Its magnitude equals the area of the parallelogram defined by the two vectors.

◇ **Exercise 8.9:** Given that $\mathbf{P}\times\mathbf{Q}=0$, what does it tell us about the vectors involved?

As an example, the vector equation of a straight line is shown below for the case where the direction of the line and one point on the line are known. Assume that $\mathbf{d}$ is a unit vector in the direction of the line and $\mathbf{P}_1$ is a given point on the line. The equation of the entire line is

$$\mathbf{P}(t)=\mathbf{P}_1+t\mathbf{d}, \tag{8.10}$$

where t can take any real value.

◇ **Exercise 8.10:** Derive the vector line equation for the straight segment between two given points $\mathbf{P}_1$ and $\mathbf{P}_2$.

- The triple product $(\mathbf{P} \bullet \mathbf{Q})\mathbf{R}$ is sometimes useful. It can be represented as

$$\begin{aligned}(\mathbf{P} \bullet \mathbf{Q})\mathbf{R} &= (P_xQ_x + P_yQ_y + P_zQ_z)(R_x, R_y, R_z) \\ &= \big((P_xQ_x + P_yQ_y + P_zQ_z)R_x, (P_xQ_x + P_yQ_y + P_zQ_z)R_y, \\ &\quad (P_xQ_x + P_yQ_y + P_zQ_z)\big)R_z \\ &= (Q_x, Q_y, Q_z)\begin{pmatrix} P_xR_x & P_yR_x & P_zR_x \\ P_xR_y & P_yR_y & P_zR_y \\ P_xR_z & P_yR_z & P_zR_z \end{pmatrix} \\ &= \mathbf{Q}(\mathbf{PR}), \end{aligned} \tag{8.11}$$

where the notation $(\mathbf{PR})$ stands for the 3×3 matrix above.

8.8.1 Direction of Rotation



Perhaps the most confusing fact about three-dimensional rotations is the direction of the rotation. The rotation is about an axis, but is it clockwise or counterclockwise? And how is this determined? The standard convention is to specify the direction a rotation according to whether it conforms to the right-hand rule or the left-hand rule, instead of using terms such as clockwise or counterclockwise. The right-hand rule says, point your right thumb in the positive direction of the axis, and curl your right fingers. They will then point in the direction of rotation as illustrated in the figure. If the fingers point in the opposite direction, then the rotation conforms to the left-hand rule. If the rotation is part of an animation, it is also common to set the length of the rotation axis so it is proportional to the speed of rotation.

Two-dimensional rotations are discussed in Section 3.6. They are simple, which is why it comes as a surprise that increasing the number of dimensions from 2 to 3 comes with an astonishing increase in complexity as well as with difficulties in visualizing the results of rotations.

> If I'm in the rotation, I'm in the rotation. If I'm not, I still will support the guys like I am in the rotation. But when I do get my turn, I definitely will make the best of it.
> —Tavon Austin.

8.8.2 Elemental Rotations

We start with the elemental three-dimensional rotations, those that are about the axes of the coordinate system. Their rotation matrices are an obvious extension of their two-dimensional counterparts. They rotate vectors by an angle θ about the coordinate axes, using the right-hand rule. The coordinate system itself should be right-handed (Figure 8.27) and a rotation is done by multiplying $R \cdot \mathbf{p}$, where the rotated point $\mathbf{p}$ is

a column vector. The elemental rotation matrices are

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix},$$
$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{8.12}$$
$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

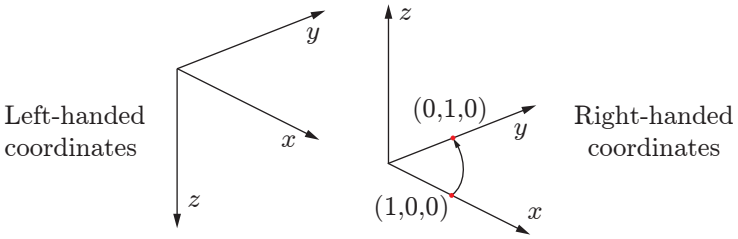


Figure 8.27: Left- and Right-Hand Coordinate Systems.

For those who prefer the terms clockwise and counterclockwise, if the rotation angle θ is positive and the coordinate system is right-handed, then the rotations caused by the matrices of Equation (8.12) would appear counterclockwise when the observer looks through the rotation axis toward the origin (and conversely, they would appear clockwise when viewed from the origin). An example is the point $(1, 0, 0)$ rotated 90° about the z-axis, as in Figure 8.27. When viewed from the positive z-axis, looking down toward the origin, this rotation is counterclockwise, but when viewed from the origin, looking up the z-axis, it seems (if you can see it) clockwise.

$$R_z(90^\circ)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

The rotation matrices of Equation (8.12) are simple but not very useful. In practice we often have to deal with general rotations, and it is difficult to break an arbitrary three-dimensional rotation into three rotations about the coordinate axes. There are some cases, however, where the basic rotations about the coordinate axes are common, and here are two examples:

- The L-systems, described in detail on Page 418.
- Rotations about the coordinate axes are common in the motion of a submarine or an airplane. These vehicles have three degrees of freedom and have three natural, mutually

perpendicular axes of rotation that are called roll, pitch, and yaw (Figure 8.28). Roll is a rotation about the direction of motion of the vehicle, or in general a rotation around its longitudinal axis. An airplane rolls when it banks by dipping one wing and lifting the other. Pitch is an up or down rotation about an axis that goes through the wings (the lateral axis). An airplane uses its elevators for this. Yaw is a left–right rotation about a vertical axis, accomplished by the rudder. These terms originated with sailors because a ship can yaw and also has (necessarily limited) roll and pitch capabilities.

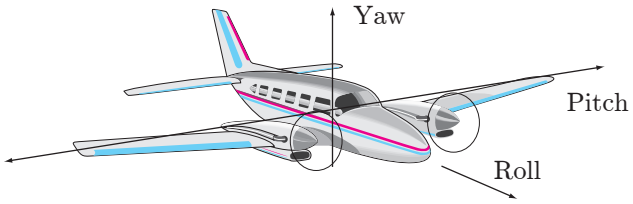


Figure 8.28: Roll, Pitch, and Yaw.

> An object is chiral if it cannot be brought into congruence with its mirror image by translation and rotation.
>
> —Vladimir Prelog.

8.8.3 General Rotations

A general rotation in three-dimensions can be achieved by a succession of three rotations, and this can be done in two ways, referred to as extrinsic and intrinsic rotations.

Extrinsic rotations rotate a point or an object around the three coordinate axes, where each rotation is by its own angle. The angles are usually the Tait-Bryan angles, also known as yaw, pitch, and roll (Figure 8.28). There are six sequences of extrinsic Tait-Bryan rotations, and they are denoted by $x-y-z$, $y-z-x$, $z-x-y$, $x-z-y$, $z-y-x$, $y-x-z$.

An intrinsic rotation may best be understood by considering a sphere, whose north pole A is located on the z-axis. Figure 8.29 shows how point B, on the equator, is initially located on the x-axis. We start by rotating the sphere about the z-axis (point A), which brings B to a point x' between the x- and y axes. We then rotate about x', which in turn moves A to a point z'' between the z- and x axes. The third and final rotation is about z''. Thus, this intrinsic rotation is about the axes of the rotating coordinate system (z, x', z''), that rotates rigidly with the moving body, and may change its orientation with respect to the (static) coordinate system after each rotation.

The three Euler angles, Figure 8.30, were introduced by Euler to describe the orientation of a rigid body with respect to a fixed coordinate system. When Euler angles are used in an intrinsic rotation, the six possible sequences of rotations are $z-x-z$, $x-y-x$, $y-z-y$, $z-y-z$, $x-z-x$, $y-x-y$.

The blue parts of Figure 8.30 indicate the original coordinate system. The red part show the rotated coordinate system, and the green line is the line of nodes, where the original xy plane and the rotated XY plane intersect. The Euler angles have the following meaning:

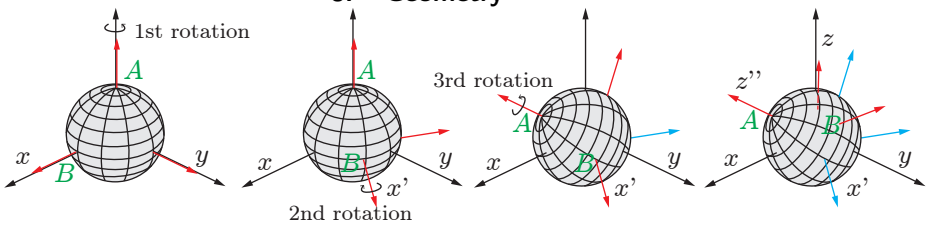
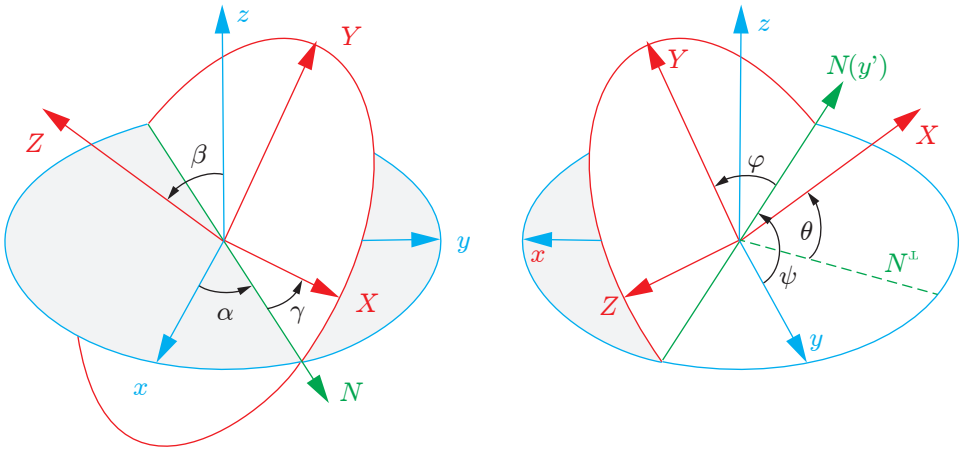


Figure 8.29: An Intrinsic Rotation.

- In the x-convention, angle α is the signed angle between the x axis and the N axis. There is also a y-convention, where α is between y and N.
- β is the angle between the z axis and the Z axis.
- γ is the signed angle between the N axis and the X axis (the x-convention).



Euler angles
Author: Lionel Brits
licensed under the Creative Commons

Tait-Bryan angles, ZYX convention
Author: Juansempere
licensed under the Creative Commons

Figure 8.30: Euler and Tait-Bryan Angles.

The Tate-Bryan angles of Figure 8.30 are for the rotation sequence $z \to y' \to x''$. The notation $N(y')$ means that N, the line of nodes, coincides with y'. The angle rotation sequence is ψ, θ, ϕ. Also, the figure corresponds to the case $\psi > 0$ and $\theta < 0$. These conventions serve to emphasize the fact rotations in three dimensions are complex and difficult to visualize.

The three basic rotation matrices of Equation (8.12) can be multiplied together to form a single matrix. We start with a matrix that employs the Euler angles to perform

an extrinsic rotation that combines yaw, pitch, and roll, in this order, with angles α, β, and γ, respectively.

$$R_z(\gamma)R_y(\beta)R_x(\alpha) = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$
$$= \begin{bmatrix} \cos\beta\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \cos\beta\sin\gamma & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma \\ -\sin\beta & \sin\alpha\cos\beta & \cos\alpha\cos\beta \end{bmatrix}. \quad (8.13)$$

Next is a rotation matrix for an intrinsic rotation that combines yaw, pitch ,and roll with Tait-Bryan angles α, β, and γ, respectively

$$R_z(\alpha)R_y(\beta)R_x(\gamma) = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{bmatrix}$$
$$= \begin{bmatrix} \cos\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma \\ -\sin\beta & \cos\beta\sin\alpha & \cos\beta\cos\gamma \end{bmatrix}. \quad (8.14)$$

8.8.4 Coordinate-Free Rotations

This section treats the problem of three-dimensional rotation with the help of vector operations and without referring to a coordinate system. Points are referred to not by their coordinates relative to the origin, but as the endpoints of vectors.

An axis is a mathematical object that specifies both a location in space and a direction. A direction in three-dimensional space is specified by a vector with three components (x, y, z), but a vector (Section 8.9) does not have a definite location in space. Its only attributes are direction and magnitude. This is why specifying an axis requires more than three numbers. An axis can be specified by the coordinates of its start (tail) and end (head) points, or by its start point and direction. Either specification requires six numbers. The three coordinate axes are an exception. We can think of them as starting at the origin, which is why three numbers are enough to specify these axes.

The case of two-dimensional rotations, Section 3.6, is simpler and may help the reader understand how three-dimensional rotations can be implemented in practice. Rotating a two-dimensional point $\mathbf{P}$ about another point $\mathbf{Q}$ is one by: (1) Applying a transformation $\mathbf{T}$ to translate both $\mathbf{P}$ and $\mathbf{Q}$ as one unit such that $\mathbf{Q}$ ends up at the origin. (2) Rotating $\mathbf{P}$ about the origin to point $\mathbf{P}^*$. (3) Applying the inverse $\mathbf{T}^{-1}$ to translate $\mathbf{P}^*$ to its correct location.

A similar procedure is performed in the case of a three-dimensional rotation. Given a vector $\mathbf{v}$ and a point $\mathbf{P}$ to rotate about it, we implicitly assume that $\mathbf{v}$ starts at the origin, we perform the rotation and determine the location of the rotated point $\mathbf{P}^*$. We then transform both $\mathbf{P}$ and $\mathbf{P}^*$ rigidly, such that $\mathbf{P}$ ends ups at its original location.

In practice, we generally don't know how to express an arbitrary rotation as a product of rotations about the coordinate axes (but see subsection 8.8.6). All we know is that a point $\mathbf{P}$ is to be rotated through an angle θ about a given axis. The previous

paragraph explains the difference between an axis and a vector, and the approach to rotations that is presented here ignores axes and concentrates on vectors.

We assume that a vector $\mathbf{r}$ is to be rotated about another (unit) vector $\mathbf{u}$ to become the rotated vector $\mathbf{r}*$. Once $\mathbf{r}*$ is obtained, it is easy to rigidly move vectors $\mathbf{r}$ and $\mathbf{r}*$ anywhere in space, and use the coordinates of their endpoints $\mathbf{P}$ and $\mathbf{P}^*$ for any further calculations and processing.

Figure 8.31a shows that the component OC of $\mathbf{r}$ along $\mathbf{u}$ is left unchanged, but the component CP is rotated to CP*. The distance OC is seen from the diagram to be $(\mathbf{r}\bullet\mathbf{u})$, so the vector $\vec{\mathrm{OC}}$ can be written $(\mathbf{r}\bullet\mathbf{u})\mathbf{u}$. From $\mathbf{r}=\vec{\mathrm{OC}}+\vec{\mathrm{CP}}$, we get $\vec{\mathrm{CP}}=\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}$ or, in terms of magnitudes, $|\vec{\mathrm{CP}}|=|\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}|$. It can also be seen from the diagram that $|\vec{\mathrm{CP}}|=|\mathbf{r}|\sin\phi$. Since $\mathbf{u}$ is a unit vector, we can write $|\mathbf{u}\times\mathbf{r}|=|\mathbf{r}|\sin\phi$. We thus obtain $|\vec{\mathrm{CP}}|=|\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}|=|\mathbf{u}\times\mathbf{r}|$.

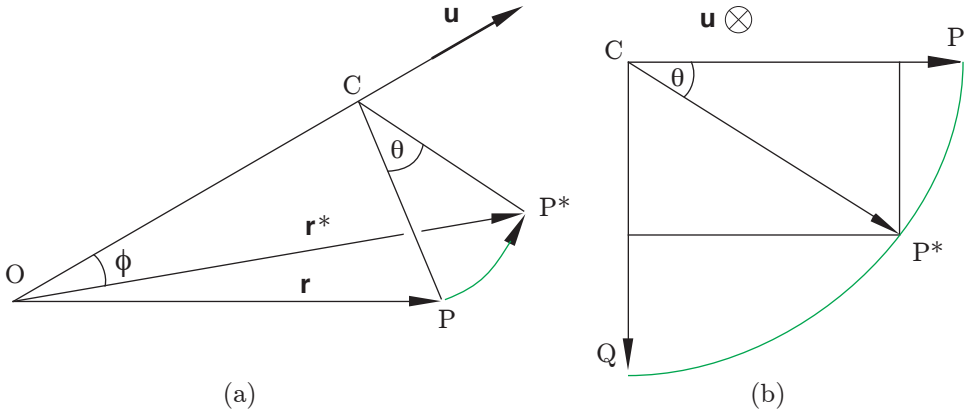


Figure 8.31: A Vector Rotation.

Figure 8.31b shows the situation when looking in the positive $\mathbf{u}$ direction. (The diagram shows the tail of $\mathbf{u}$.) Note that the vector $\vec{\mathrm{CQ}}$ is perpendicular to both $\mathbf{u}$ and $\mathbf{r}$, so it is in the direction of $\mathbf{u}\times\mathbf{r}$.

The next step is to resolve CP* into its components. From Figure 8.31b, we get

$$\vec{\mathrm{CP}}^*=\cos\theta[\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}]+\sin\theta[\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}]=\cos\theta[\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}]+\sin\theta(\mathbf{u}\times\mathbf{r}),$$

which can be used to express $\mathbf{r}^*$:

$$\mathbf{r}^*=\vec{\mathrm{OC}}+\vec{\mathrm{CP}}^*=(\mathbf{r}\bullet\mathbf{u})\mathbf{u}+\cos\theta[\mathbf{r}-(\mathbf{r}\bullet\mathbf{u})\mathbf{u}]+\sin\theta(\mathbf{u}\times\mathbf{r}). \tag{8.15}$$

Using Equations (8.9) and (8.11) on Page 453, we can rewrite this as $\mathbf{r}^*=(\mathbf{u}\mathbf{u}^T)\mathbf{r}+\cos\theta\mathbf{r}-\cos\theta(\mathbf{u}\mathbf{u}^T)\mathbf{r}+\sin\theta\mathbf{U}\mathbf{r}$, where

$$\mathbf{U}=\begin{pmatrix}0 & -u_z & u_y\\ u_z & 0 & -u_x\\ -u_y & u_x & 0\end{pmatrix}.$$

The result can now be summarized as $\mathbf{r}^* = \mathbf{Mr}$, where

$$\mathbf{M} = \mathbf{uu}^T + \cos\theta(\mathbf{I} - \mathbf{uu}^T) + \sin\theta\mathbf{U} \tag{8.16}$$
$$= \begin{bmatrix} u_x^2 + \cos\theta(1-u_x^2) & u_xu_y(1-\cos\theta) - u_z\sin\theta & u_xu_z(1-\cos\theta) + u_y\sin\theta \\ u_xu_y(1-\cos\theta) + u_z\sin\theta & u_y^2 + \cos\theta(1-u_y^2) & u_yu_z(1-\cos\theta) - u_x\sin\theta \\ u_xu_z(1-\cos\theta) - u_y\sin\theta & u_yu_z(1-\cos\theta) + u_x\sin\theta & u_z^2 + \cos\theta(1-u_z^2) \end{bmatrix}.$$

Direction cosines. If $\mathbf{v} = (v_x, v_y, v_z)$ is a three-dimensional vector, its *direction cosines* are defined as

$$N_1 = \frac{v_x}{\| \mathbf{v} \|}, \quad N_2 = \frac{v_y}{\| \mathbf{v} \|}, \quad N_3 = \frac{v_z}{\| \mathbf{v} \|}.$$

These are the cosines of the angles between the direction of $\mathbf{v}$ and the three coordinate axes. It is easy to verify that $N_1^2 + N_2^2 + N_3^2 = 1$. If $\mathbf{u} = (u_x, u_y, u_z)$ is a unit vector, then $\| \mathbf{u} \| = 1$ and u_x, u_y, and u_z are the direction cosines of $\mathbf{u}$.

It can be shown that a rotation through an angle $-\theta$ is performed by the transpose $\mathbf{M}^T$. Consider the two successive and opposite rotations $\mathbf{r}^* = \mathbf{Mr}$ and $\mathbf{r}' = \mathbf{M}^T\mathbf{r}^*$. On the one hand, they can be expressed as the product $\mathbf{r}' = \mathbf{M}^T\mathbf{r}^* = \mathbf{M}^T\mathbf{Mr}$. On the other hand, they rotate in opposite directions, so they return all points to their original positions; therefore $\mathbf{r}'$ must be equal to $\mathbf{r}$. We end up with $\mathbf{r} = \mathbf{M}^T\mathbf{Mr}$ or $\mathbf{MM}^T = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The transpose $\mathbf{M}^T$ therefore equals the inverse, $\mathbf{M}^{-1}$, of $\mathbf{M}$, which shows that a rotation matrix $\mathbf{M}$ is orthogonal.

Example: Consider a rotation about the z axis. The rotation axis is $\mathbf{u} = (0, 0, 1)$, resulting in

$$\mathbf{uu}^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{U} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ and hence } \mathbf{M} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which is the familiar rotation matrix about the z axis. It is identical to the z-rotation matrix of Equation (8.12), so we conclude that it rotates counterclockwise when viewed from the z axis, looking to the origin.

The general rotation matrix of Equation (8.16) can also be constructed as the product of five simple rotations about various coordinate axes. Given a unit vector $\mathbf{u} = (u_x, u_y, u_z)$, consider the following rotations.

1. Rotate $\mathbf{u}$ about the z axis into the xz plane, so its y coordinate becomes zero. This is done by a rotation matrix of the form

$$\mathbf{A} = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the angle ψ of rotation can be computed from the requirement that the y component of vector $\mathbf{v} = \mathbf{uA}$ be zero. This component is $-u_x\sin\psi + u_y\cos\psi$, which implies $\cos\psi = u_x/\sqrt{u_x^2 + u_y^2}$ and $\sin\psi = u_y/\sqrt{u_x^2 + u_y^2}$. Notice that rotating $\mathbf{u}$ does not affect

its magnitude, so $\mathbf{v}$ is also a unit vector. In addition, since the rotation is about the z axis, the z component of $\mathbf{u}$ does not change, so $v_z = u_z$.

2. Rotate vector $\mathbf{v}$ about the y axis until it coincides with the z axis. This is accomplished by the matrix

$$\mathbf{B} = \begin{bmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{bmatrix}.$$

The angle ϕ of rotation is computed from the dot product $\cos\phi = \mathbf{v}\cdot(0,0,1) = v_z = u_z$, implying that $\sin\phi = \sqrt{1-u_z^2}$. Since $\mathbf{v}$ is a unit vector, it is rotated by $\mathbf{B}$ to vector $(0,0,1)$.

3. Rotate $(0,0,1)$ about the z axis through an angle θ. This is done by matrix

$$\mathbf{C} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This is a trivial rotation that does not change $(0,0,1)$.

4. Rotate the result of step 3 by $\mathbf{B}^{-1}$ (which equals $\mathbf{B}^T$).

5. Rotate the result of step 4 by $\mathbf{A}^{-1}$ (which equals $\mathbf{A}^T$).

When these five steps are performed on a point (x,y,z), the effect is to rotate the point through an angle θ about $\mathbf{u}$. In practice, the five steps are combined by multiplying the five matrices above, as shown in the listing of Figure 8.32. The result is identical to Equation (8.16).

```
tm=Sqrt[x^2+y^2];
a={{x/tm,-y/tm,0},{y/tm,x/tm,0},{0,0,1}};
b={{z,0,Sqrt[1-z^2]},{0,1,0},{-Sqrt[1-z^2],0,z}};
c={{Cos[t],-Sin[t],0},{Sin[t],Cos[t],0},{0,0,1}};
FullSimplify[a.b.c.Transpose[b].Transpose[a] /. x^2+y^2->1-z^2]
```

Figure 8.32: Mathematica Code for a General Rotation.

8.8.5 Rodrigues' Rotation Formula

The Rodrigues' rotation formula [WikiRodrigues 25] is another coordinate-free approach to three-dimensional rotations. It is similar to the discussion in the previous subsection. Instead of relying on three coordinate axes and an origin, vector analysis is used to derive a formula for rotating a vector $\mathbf{v}$ through an angle θ about a (unit) vector $\mathbf{k}$. Figure 8.33 illustrates the relevant vectors used in deriving the Rodrigues rotation formula. $\mathbf{k}$ is a unit vector about which vector $\mathbf{v}$ is rotated by an angle θ. The rotation complies with the right-hand rule.

The first step in deriving the Rodrigues rotation formula is to express the rotated vector $\mathbf{v}$ as the vector sum $\mathbf{v}_\parallel + \mathbf{v}_\perp$, where $\mathbf{v}_\parallel = (\mathbf{k}\cdot\mathbf{v})\mathbf{k}$ and $\mathbf{v}_\perp = \mathbf{v} - \mathbf{v}_\parallel$.

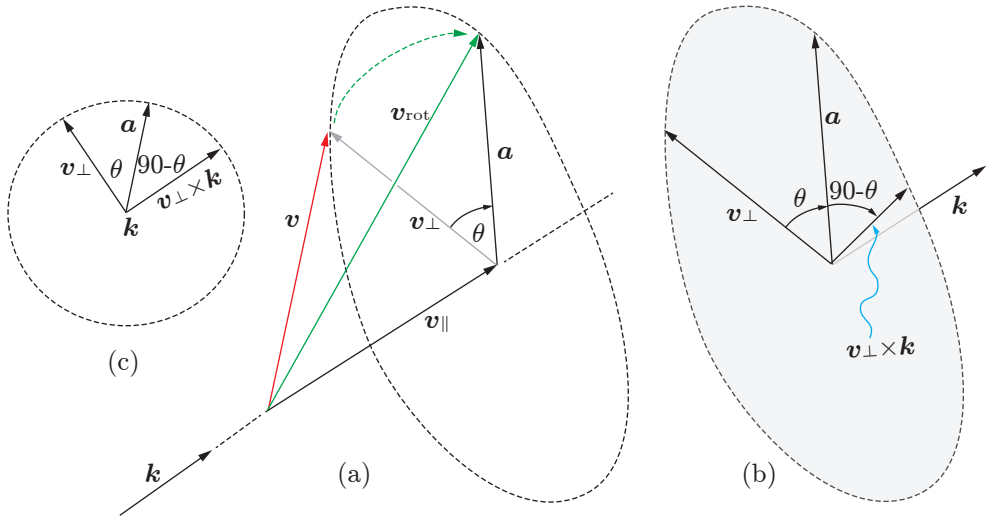


Figure 8.33: Construction for the Rodrigues' Rotation Formula.

A glance at Figure 8.33a shows that the rotated vector is the vector sum $\mathbf{v}_{rot} = \mathbf{v}_{\parallel} + a$, which leaves vector $\mathbf{a}$ as the only remaining unknown. Parts (b) and (c) of the figure attempt to show that this vector is the vector sum of $\mathbf{v}_{\perp}\cos\theta$ and $\mathbf{v}_{\perp} \times \mathbf{k}\sin\theta$. Figure 8.33b illustrates how vectors $\mathbf{v}_{\perp}$ and $\mathbf{a}$ are both located in the plane of rotation, which is perpendicular to $\mathbf{k}$. Figure 8.33c attempts to illustrate the sum above by showing that vector $\mathbf{v}_{\perp} \times \mathbf{k}$ is also resides in the same plane and is perpendicular to $\mathbf{v}_{\perp}$. This part of the figure also shows that the cross-product $\mathbf{v}_{\perp} \times \mathbf{k}$ conforms to the right-hand rule. Those who are interested in a left-hand rotation need only to reverse the order of the components of this cross product.

The final result, the Rodrigues rotation formula, is therefore

$$\begin{aligned}\mathbf{v}_{rot} &= \mathbf{v}_{\parallel} + \mathbf{v}_{\perp}\cos\theta + \mathbf{v}_{\perp} \times k\sin\theta \\ &= (\mathbf{k}\cdot\mathbf{v})\mathbf{k} + [\mathbf{v} - (\mathbf{k}\cdot\mathbf{v})\mathbf{k}]\cos\theta + [\mathbf{v} - (\mathbf{k}\cdot\mathbf{v})\mathbf{k}] \times \mathbf{k}\sin\theta \\ &= (\mathbf{k}\cdot\mathbf{v})\mathbf{k}[1 - \cos\theta] + \mathbf{v}\cos\theta + \mathbf{v} \times \mathbf{k}\sin\theta. \end{aligned} \tag{8.17}$$

Notice that the term $(\mathbf{k}\cdot\mathbf{v})\mathbf{k} \times \mathbf{k}\sin\theta$ is zero, because it is the cross-product of two parallel vectors.

> Benjamin Olinde Rodrigues, more commonly known as Olinde Rodrigues, was a French banker, mathematician, and social reformer. In mathematics Rodrigues is remembered for Rodrigues' rotation formula for vectors, the Rodrigues formula for the Legendre polynomials, and the Euler–Rodrigues parameters.
>
> —From Wikipedia.

8.8.6 Givens Rotations

The general rotation matrix, Equation (8.16), can be constructed for any general rotation in three dimensions. Given such a matrix $\mathbf{A}$, it is possible to reduce it to a product of rotation matrices that cause the same rotation by performing a sequence of rotations about the coordinate axes. This process, first described in [Givens 58], is based on the QR decomposition of matrices, a subject discussed in any text on matrices, and it results in a set of *Givens rotations*. Each Givens rotation matrix $\mathbf{T}_{i,j}$ is identified by two indexes, i and j, where $i > j$. The matrix is an identity matrix except for the two diagonal elements (i,i) and (j,j) that are cosines of some angle and for the two off-diagonal elements (i,j) and (j,i) that are the $\pm\sin$ of the same angle. Specifically, $\mathbf{T}_{i,j}[i,i] = \mathbf{T}_{i,j}[j,j] = c$ and $\mathbf{T}_{i,j}[j,i] = -\mathbf{T}_{i,j}[i,j] = s$, where $c = \mathbf{A}[j,j]/D$, $s = \mathbf{A}[i,j]/D$, and $D = \sqrt{\mathbf{A}[j,j]^2 + \mathbf{A}[i,j]^2}$. The special construction of $\mathbf{T}_{i,j}$ implies that the matrix product $\mathbf{T}_{i,j}\mathbf{A}$ transforms $\mathbf{A}$ to a matrix whose (i,j)th element is zero.

Once a general rotation matrix $\mathbf{A}$ is given, its Givens rotations can be found by preparing the Givens rotation matrices $\mathbf{T}_{i,j}$ that zero those elements of $\mathbf{A}$ located below the main diagonal, column by column, from the bottom up. Figure 8.34 is a listing of Matlab code that does that for the rotation matrix that rotates point $(1,1,1)$ to the x axis.

```
n=3;
A=[.5774,-.5774,-.5774; .5774,.7886,-.2115; .5774,-.2115,.7886]
% Rotation from 1,1,1 to x-axis
Q=eye(n);
for j=1:n-1,
  for i=n:-1:j+1,
    T=eye(n);
    D=sqrt(A(j,j)^2+A(i,j)^2);
    cos=A(j,j)/D; sin=A(i,j)/D;
    T(j,j)=cos; T(j,i)=sin; T(i,j)=-sin; T(i,i)=cos; T
    A=T*A;
Q=Q*T';
  end;
end;
Q
A
```

Figure 8.34: Computing Three Givens Matrices.

The three rotation matrices produced by this computation are listed in Figure 8.35, where they are used to rotate point $(1,1,1)$ to the x axis. Matrix `T1` rotates $(1,1,1)$ 45° about the y axis to $(1.4142,1,0)$, which is rotated by `T2` 35.26° about the z axis to $(1.7321,0,0)$, which is trivially rotated by `T3` 15° about the x axis to itself.

8.8.7 Fast Rotations

Rotation requires the calculation of the transcendental sine and cosine functions, which is time consuming. If many rotations are needed, it is preferable to precompute the trigonometric functions for many angles and store them in a table. This section shows

```
T1=[0.7071,0,0.7071; 0,1,0; -0.7071,0,0.7071];
T2=[0.8165,0.5774,0; -0.5774,0.8165,0; 0,0,1];
T3=[1,0,0; 0,0.9660,0.2587; 0,-0.2587,0.9660];
p=[1;1;1];
a=T1*p
b=T2*a
c=T3*b
```

Figure 8.35: Rotating Point $(1, 1, 1)$ to the x Axis.

J. Wallace Givens, Jr. (1910–1993) pioneered the use of plane rotations in the early days of automatic matrix computations. Givens graduated from Lynchburg College in 1928, and he completed his Ph.D. at Princeton University in 1936. After spending three years at the Institute for Advanced Study in Princeton as an assistant of Oswald Veblen, Givens accepted an appointment at Cornell University, but later moved to Northwestern University. In addition to his academic career, Givens was the director of the Applied Mathematics Division at Argonne National Lab and, like his counterpart Alston Householder at Oak Ridge National Laboratory, Givens served as an early president of SIAM. He published his work on the rotations in 1958.

—Carl D. Meyer.

how to do this using only integers, a method that results in much faster rotations than using floating-point numbers.

The method is illustrated for the first quadrant (rotation angles of 0° to 90°) in increments of 1°. Notice that rotations in other quadrants can be achieved by a first-quadrant rotation followed by a reflection. The following Mathematica code generates 91 sine values, from $\sin 0° = 0$ to $\sin 90° = 1$, multiplies each by $2^{14} = 16{,}384$, rounds them, and stores them in a table as 16-bit integers ranging from 0 to 16,384.

```
d2r=Pi/180;
Table[Round[N[16384*Sin[i*d2r]]], {i,0,90}]
```

The 91 values are listed in Table 8.36, but notice that they are only approximations of the true sine values. (Even floating-point sine values are, in general, just approximations, but normally better than our integers.) This means that the use of this table for many successive rotations of a point may place it farther and farther away from its true position. When we perform many successive rotations of an object that consists of many points, placing points away from where they should be generally results in a deformation of the object.

We assume that the points are represented by coordinates that are 16-bit integers. Calculating the rotated coordinates (x^*, y^*) of a point (x, y) can now be done, for example, by

$$x^* = \text{rshift}(x \times \text{Table}(90 - \theta), 14) - \text{rshift}(y \times \text{Table}(\theta), 14),$$
$$y^* = \text{rshift}(x \times \text{Table}(\theta), 14) + \text{rshift}(y \times \text{Table}(90 - \theta), 14).$$

θ	$\sin\theta$	θ	$\sin\theta$	θ	$\sin\theta$	θ	$\sin\theta$	θ	$\sin\theta$
0	0	1	286	2	572	3	857	4	1143
5	1428	6	1713	7	1997	8	2280	9	2563
10	2845	11	3126	12	3406	13	3686	14	3964
15	4240	16	4516	17	4790	18	5063	19	5334
20	5604	21	5872	22	6138	23	6402	24	6664
25	6924	26	7182	27	7438	28	7692	29	7943
30	8192	31	8438	32	8682	33	8923	34	9162
35	9397	36	9630	37	9860	38	10087	39	10311
40	10531	41	10749	42	10963	43	11174	44	11381
45	11585	46	11786	47	11982	48	12176	49	12365
50	12551	51	12733	52	12911	53	13085	54	13255
55	13421	56	13583	57	13741	58	13894	59	14044
60	14189	61	14330	62	14466	63	14598	64	14726
65	14849	66	14968	67	15082	68	15191	69	15296
70	15396	71	15491	72	15582	73	15668	74	15749
75	15826	76	15897	77	15964	78	16026	79	16083
80	16135	81	16182	82	16225	83	16262	84	16294
85	16322	86	16344	87	16362	88	16374	89	16382
90	16384								

Table 8.36: Sine Values as 16-Bit Integers.

Notice how the required cosine values are obtained from the end of the table. This method works because the table has 91 entries. Multiplying a 16-bit integer coordinate by a 16-bit integer sine value yields a 32-bit product. The right shift effectively divides the product by $2^{14} = 16{,}384$, a necessary operation because our integer sine values have been premultiplied by this scale factor.

◇ **Exercise 8.11:** Use this method to calculate the results of rotating point $(1, 2)$ by $60°$ and by $80°$. In each case, compare the results with those obtained when built-in sine and cosine functions are used.

8.8.8 CORDIC Rotations

We routinely use calculators to compute values of common functions, but have you ever wondered how a calculator determines the value of, say, $\tan 72.81°$ so fast? Many calculators use CORDIC (COordinate Rotation, DIgital Computer), a general method for computing many elementary functions. CORDIC was originally proposed by [Volder 59] and was extended by [Walther 71]. The original references are hard to find but are included in [Swartzlander 90]. Here, we show how CORDIC can be used to implement fast rotations.

> CORDIC, short for coordinate rotation digital computer, is a simple and efficient algorithm to calculate trigonometric functions, hyperbolic functions, square roots, multiplications, divisions, and exponentials and logarithms with arbitrary base, typically converging with one digit (or bit) per iteration. CORDIC is therefore an example of a digit-by-digit algorithm. The original system is sometimes referred to as Volder's algorithm.
>
> —From `https://en.wikipedia.org/wiki/CORDIC`

It is sufficient to consider a rotation about the origin where the rotation angle θ is in the interval $[0, 90°)$ (the first quadrant). The special case $\theta = 90°$ can be implemented by the negate and exchange rule

$$\begin{pmatrix} \cos(90) & -\sin(90) \\ \sin(90) & \cos(90) \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Rotations in other quadrants can be achieved by a first-quadrant rotation, followed by a reflection.

The rotation is expressed by [see Equation (3.2)]

$$(x^*, y^*) = (x, y) \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{8.18}$$

Because θ is less than 90°, we know that $\cos\theta$ is nonzero, so we can factor out $\cos\theta$, yielding

$$(x^*, y^*) = \cos\theta\,(x, y) \begin{pmatrix} 1 & -\tan\theta \\ \tan\theta & 1 \end{pmatrix}.$$

We now express θ as the sum $\sum_{i=0}^{m} \theta_i$, where angles θ_i are defined by the relation $\tan\theta_i = 2^{-i}$ or $\theta_i \stackrel{\text{def}}{=} \arctan(2^{-i})$. The first 16 θ_i, for $i = 0, 1, \ldots, 15$, are listed in Table 8.37.

In order to express any angle θ as the sum of these particular θ_i, some θ_i will have to be subtracted. Consider, for example, $\theta = 58°$. We start with $\theta_0 = 45°$. Since $\theta_0 < \theta$, we add θ_1. The sum $\theta_0 + \theta_1 = 45 + 26.5651 = 71.5651$ is greater than θ, so we subtract θ_2. The new sum, 57.5289, is less than θ, so we add θ_3, and so on.

◇ **Exercise 8.12:** We want to be able to express any angle θ in the range $[0°, 90°)$ by adding and subtracting a number of consecutive θ_i, from θ_0 to some θ_m, without skipping any θ_i in between. Is that possible?

It is easy to write a program that decides which of the θ_i's should be added and which should be subtracted. Thus, we end up with

$$\theta = \sum_{i=0}^{m} d_i\theta_i = \sum_{i=0}^{m} d_i \arctan(2^{-i}), \quad \text{where} \quad d_i = \pm 1.$$

Once the number m of necessary d_i's and their values have been determined, we rotate (x, y) to (x^*, y^*) in a loop where each iteration rotates a point (x_i, y_i) through an

i	θ_i (degrees)	θ_i(radians)	K_i
0	45.	0.785398	0.70710678118654746
1	26.5651	0.463648	0.63245553203367577
2	14.0362	0.244979	0.61357199107789628
3	7.12502	0.124355	0.60883391251775243
4	3.57633	0.0624188	0.60764825625616825
5	1.78991	0.0312398	0.60735177014129604
6	0.895174	0.0156237	0.60727764409352614
7	0.447614	0.00781234	0.60725911229889284
8	0.223811	0.00390623	0.60725447933256249
9	0.111906	0.00195312	0.60725332108987529
10	0.0559529	0.000976562	0.60725303152913446
11	0.0279765	0.000488281	0.60725295913894495
12	0.0139882	0.000244141	0.60725294104139727
13	0.00699411	0.00012207	0.60725293651701029
14	0.00349706	0.0000610352	0.60725293538591352
15	0.00174853	0.0000305176	0.60725293510313938

Table 8.37: The First 16 θ_i's and Scale Factors.

angle $d_i\theta_i$ to a point (x_{i+1}, y_{i+1}). A general iteration can be expressed in the form

$$\begin{aligned}(x_{i+1}, y_{i+1}) &= \cos(d_i\theta_i)\,(x_i, y_i)\begin{pmatrix} 1 & -d_i\tan\theta_i \\ d_i\tan\theta_i & 1\end{pmatrix}\\ &= \cos(d_i\theta_i)\,(x_i, y_i)\begin{pmatrix} 1 & -d_i2^{-i} \\ d_i2^{-i} & 1\end{pmatrix}\\ &= \cos(d_i\theta_i)\,(x_i + y_i d_i 2^{-i}, y_i - x_i d_i 2^{-i}).\end{aligned} \tag{8.19}$$

We interpret the result (x_{i+1}, y_{i+1}) of an iteration as the vector from the origin to point (x_{i+1}, y_{i+1}). Equation (8.19) shows that this vector is the product of two terms. The second term, $(x_i + y_i d_i 2^{-i}, y_i - x_i d_i 2^{-i})$, determines the direction of the vector, while the first term, $\cos(d_i\theta_i)$, affects only the magnitude of the vector. The second term is easy to calculate since it just involves shifts. We know that d_i is just a sign and that a product of the form $x_i 2^{-i}$ can be computed by shifting x_i i positions to the right. The problem is to calculate the first term, $\cos(d_i\theta_i)$, and to multiply the two terms.

This is why CORDIC proceeds by first performing all the iterations

$$(x_{i+1}, y_{i+1}) \leftarrow (x_i + y_i d_i 2^{-i}, y_i - x_i d_i 2^{-i})$$

using just right shifts and additions/subtractions; the cosine terms are ignored. The result is a vector that points in the right direction but is too long (Figure 8.38). To bring this vector to its correct size, it should be multiplied by the scale factor

$$K_m = \prod_{i=0}^{m} \cos\theta_i.$$

(Notice that $\cos(d_i\theta_i) = \cos\theta_i$ since cosine is an even function.) This is discouraging because it suggests that m multiplications are needed just to calculate the scale factor K_m. However, the first 16 scale factors are listed in Table 8.37 and even a quick glance shows that they converge to the number 0.60725.... Reference [Vachss 87] shows that K_m can be obtained simply by using the m most significant bits of this number and ignoring the rest.

Using the identity $\sin^2\theta + \cos^2\theta = 1$ and the definition $\tan\theta_i = 2^{-i}$, we get

$$\cos\theta_i = \frac{1}{\sqrt{1+\tan^2\theta_i}} = \frac{1}{\sqrt{1+2^{-2i}}},$$

which is why the scale factors of Table 8.37 were so easily calculated to a high precision by the code

```
N[Table[Product[(2^(-2i)+1)^(-1/2),{i,0,n}],{n,0,16}],17]//TableForm.
```

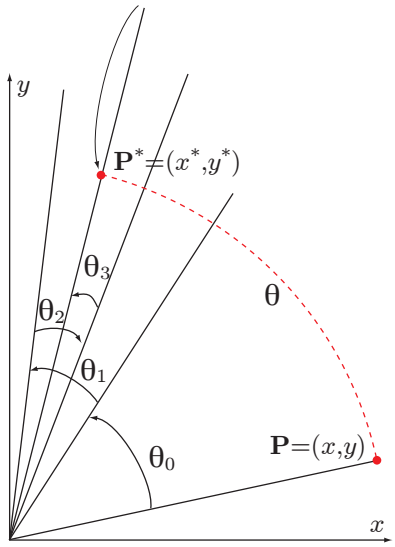


Figure 8.38: CORDIC Rotation.

◇ **Exercise 8.13:** Suggest another way to calculate K_m.

Any practical CORDIC implementation (see [Jarvis 90] for a C program) should have the following two features.

1. CORDIC employs only shifts and additions/subtractions, so any implementation should use fixed-point arithmetic instead of floating-point arithmetic. This is fast since shifting and adding fixed-point numbers can be done with integer operations. Notice that all the numbers involved in the computations are less than unity, except perhaps the original coordinates (x, y). A software package for graphics employing this method should therefore use normalized coordinates (fixed-point numbers in the interval $[0, 1]$) throughout and perform all the calculations on these small numbers. Each

iteration results in a pair (x_{i+1}, y_{i+1}) that's slightly larger than its predecessor, but the last iteration results in a pair that can be larger than (x, y) by a factor of at most $1/0.60725\ldots = 1.64676\ldots$. This pair is then scaled down when multiplied by K_m. The final step is to scale the final coordinates up.

All this suggests a 32-bit fixed-point format where the leftmost bit is reserved, as usual, for the sign, the next two bits are the integer part, and the remaining 29 bits are the fractional part (29 bits being equivalent to 9 decimal digits). The largest number that can be represented by this format is $11.11\ldots1_2 = 3.999\ldots$ and the smallest one is $110\ldots0_2 = -4$. It's a good idea to reserve two bits for the integer part because (1) even though all the numbers involved are 1 or smaller, some intermediate results may be greater than 1, and (2) this convention makes it possible to represent the important constants π, e, and ϕ (the Golden Ratio).

2. Earlier, we said, "It is easy to write a program that decides which of the θ_i's should be added and which should be subtracted." The practical way to do this is to initialize a variable z to θ and try to drive z to zero during the iterations. In iteration i the program should calculate both $z + \theta_i$ and $z - \theta_i$, select the value that's closer to zero, use it to decide whether to add or subtract θ_i, and then update z. If $z - \theta_i$ is closer to zero, then θ_i should be added; otherwise, θ_i should be subtracted. An example is $\theta = 58°$. We initialize z to 58. In iteration 0, it is clear that $58 - 45 = 13$ is closer to zero than $58 + 45$. The program therefore adds θ_0 and updates z to 13. In iteration 1, the program finds that $13 - 26.5651 = -13.5651$ is closer to zero than $13 + 26.5651$, so it adds θ_1 and updates z to -13.5651. In iteration 2, the program discovers that $-13.5651 + 14.0362 = 0.4711$ is closer to zero than $-13.5651 - 14.0362$, so it subtracts θ_2 and updates z to 0.4711.

Finally, we realize that there is really no need to compare $z+\theta_i$ and $z-\theta_i$ in iteration i. We simply start by selecting $d_0 = +1$ and update z by subtracting $z \leftarrow z - \theta_0$, $z \leftarrow z - \theta_1$, etc., until we get a negative value in z. We then change d_i to -1 (the new sign of z) and update z by $z \leftarrow z - d_i\theta_i$ (which now amounts to adding θ_i to z). This is summarized by the Mathematica code of Figure 8.39. (But note that the `Sign` function of Mathematica returns $+1$, 0, or -1, while we need a result of $+1$ or -1. The code as shown is simple but not completely general.)

```
t=Table[ArcTan[2.^{-i}], {i,0,15}]; (* arctans in radians *)
d=1; x=2.1; y=0.34; z=46. Degree;
Do[{Print[i,", ",x,", ",y,", ",z,", ",d],
 xn=x+y d 2^{-i}, yn=y-x d 2^{-i},
 zn=z-d t[[i+1]], d=Sign[zn], x=xn, y=yn, z=zn}, {i,0,14}]
Print[0.60725x,", ",0.60725y]
```

Figure 8.39: Mathematica Code for CORDIC Rotations.

> Compared to other approaches, CORDIC is a clear winner when a hardware multiplier is unavailable (e.g. in a microcontroller) or when you want to save the gates required to implement one (e.g. in an FPGA). On the other hand, when a hardware multiplier is available (e.g. in a DSP microprocessor), table-lookup methods and good old-fashioned power series are generally faster than CORDIC.
>
> —Grant R. Griffin, `www.dspguru.com/info/faqs/cordic.htm`

◇ **Exercise 8.14:** Instead of using the complex CORDIC method, wouldn't it be simpler to perform a rotation by a direct use of Equation (8.18)? After all, this only requires the calculation of one sine and one cosine values.

8.9 Points and vectors

Students frequently become bewildered by points and vectors as the mathematical notations for both entities are quite comparable. Euclid defined a point $\mathbf{P} = (x, y)$ as a geometric entity of zero dimensions. Its only attribute is its position, represented by its x and y coordinates. On the other hand, a vector is a mathematical entity with two characteristics, direction and magnitude; it has no specific position in space. Therefore, the direction of the vector $\mathbf{v} = (a, b)$ is established by the procedure: Begin at a random point $\mathbf{A}$ and proceed a units in the x direction and b units in the y direction, arriving at point $\mathbf{B}$. The orientation of vector $\mathbf{v}$ indicates the path from $\mathbf{A}$ to $\mathbf{B}$. The magnitude of $\mathbf{v}$ is represented as $\sqrt{a^2 + b^2}$.

Points and vectors may exist in spaces of other dimensions, and the study of vector spaces, of finite or infinite dimensions, is part of modern algebra. Page 113 has more to say about the relation of points to complex numbers.

In this text, points are denoted by uppercase boldface letters and may also have indexes. For example $\mathbf{A}$, $\mathbf{B}$, $\mathbf{P}$, and $\mathbf{P}_i$. Vectors are denoted by lowercase boldface letters. The notation $\mathbf{P}(t)$ refers to a point that is a function of a real parameter t.

Mathematical objects and entities are useful and interesting because they can be operated on. Entities without operations are static and have limited applications. This also applies to points and vectors. We start with operations on points. The product $\alpha\mathbf{P}$ of a real number and a point is well defined and is a point on the straight, infinite line connecting $\mathbf{P}$ to the origin. Its exact location on this line depends on α. (Notice that α can also be negative.) Subtracting points is also a simple, well-defined operation, and it produces a vector. Given points $\mathbf{P}_0 = (x_0, y_0)$ and $\mathbf{P}_1 = (x_1, y_1)$, their difference $\mathbf{P}_1 - \mathbf{P}_0$ is the pair $(x_1 - x_0, y_1 - y_0)$ and it is simply the vector from $\mathbf{P}_0$ to $\mathbf{P}_1$.

Example: The differences $(30, 45) - (17, 60) = (13, -15)$ and $(25, 40) - (12, 55) = (13, -15)$ are the same. They specify a vector whose magnitude is $\sqrt{13^2 + (-15)^2}$ and whose direction is 15 units in the negative y direction for every 13 units in the positive x direction.

The sum $\mathbf{P} + \mathbf{v}$ of a point and a vector is well defined. It is the point $\mathbf{P}$ shifted in the direction and by the magnitude of vector $\mathbf{v}$. Once this is clear, it is obvious that the expression $\mathbf{P}_0 + \alpha(\mathbf{P}_2 - \mathbf{P}_0)$ is the sum of a point and a vector, which makes it a point. We denote this point by $\mathbf{P}_1$ and observe that (1) the vector $(\mathbf{P}_2 - \mathbf{P}_0)$ points

from $\mathbf{P}_0$ to $\mathbf{P}_2$, and (2) adding this vector to $\mathbf{P}_0$ produces a point ($\mathbf{P}_1$) located on the line connecting $\mathbf{P}_0$ to $\mathbf{P}_2$. Hence, the three points $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ are collinear.

The next step is to write

$$\mathbf{P}_1 = \mathbf{P}_0 + \alpha(\mathbf{P}_2 - \mathbf{P}_0) = (1-\alpha)\mathbf{P}_0 + \alpha\mathbf{P}_2, \tag{8.20}$$

which implies that (1) point $\mathbf{P}_1$ is a combination of $\mathbf{P}_0$ and $\mathbf{P}_2$, and (2) the combination is linear. In general, each of three collinear points can be expressed as a linear combination of the other two, thereby indicating that such points are not independent. Figure 8.40 illustrates how point $\mathbf{P}$ slides on the line that passes through $\mathbf{A}$ and $\mathbf{C}$. For negative t values, $\mathbf{P}$ is beyond $\mathbf{C}$. For values $t \in [0, 1]$, point $\mathbf{P}$ slides from $\mathbf{C}$ to $\mathbf{A}$, and for positive t, point $\mathbf{P}$ is located beyond $\mathbf{A}$.

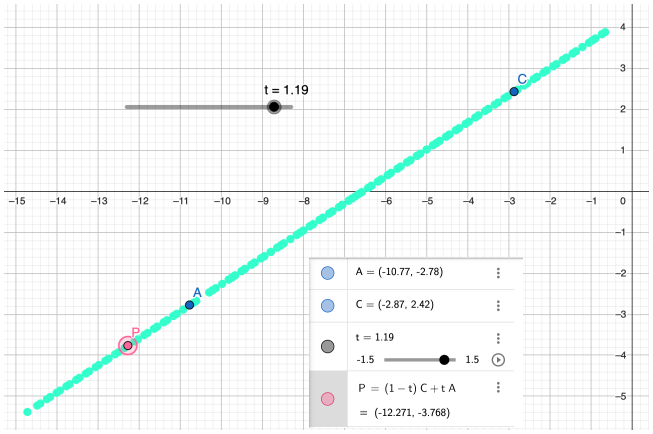


Figure 8.40: Collinear Points.

◇ **Exercise 8.15:** What can we say about four collinear points?

In contrast to the subtraction of points, adding points is in general meaningless, as the following shows. Simply adding points $(0, 1) + (6, 11)$, produces $(6, 12)$. However, when the points are shifted and then added, the new sum is not shifted by the same amount. Shifting by 5 results in the sum $(0 + 5, 1) + (6 + 5, 11) = (16, 12)$, which is shifted by 10.

There is, however, an important special case where the sum of points is well defined, the so-called *barycentric sum*. Given a set of points, we multiply each by a weight (a real number). If the weights add up to 1, then the sum of the weighted points is well defined in the sense that shifting the points rigidly (all by the same amount) would shift the sum by the same amount. We say that a barycentric sum is affinely invariant. Here is the (simple) proof: If $\sum_{i=0}^{n} w_i = 1$, then

$$\sum_{i=0}^{n} w_i\mathbf{P}_i = \mathbf{P}_0 + \sum_{i=1}^{n} w_i\mathbf{P}_i - (1-w_0)\mathbf{P}_0$$

$$
\begin{aligned}
&= \mathbf{P}_0 + w_1\mathbf{P}_1 + w_2\mathbf{P}_2 + \cdots + w_n\mathbf{P}_n - (w_1 + \cdots + w_n)\mathbf{P}_0 \\
&= \mathbf{P}_0 + w_1(\mathbf{P}_1 - \mathbf{P}_0) + w_2(\mathbf{P}_2 - \mathbf{P}_0) + \cdots + w_n(\mathbf{P}_n - \mathbf{P}_0) \\
&= \mathbf{P}_0 + \sum_{i=1}^{n} w_i(\mathbf{P}_i - \mathbf{P}_0). \qquad (8.21)
\end{aligned}
$$

This is the sum of the point $\mathbf{P}_0$ and the vector $\sum_{i=1}^{n} w_i(\mathbf{P}_i - \mathbf{P}_0)$, and we already know that the sum of a point and a vector is a point.

Barycentric sums are common in computer graphics, especially in curve and surface design. Many methods for curves and surfaces are based on weighted sums of points, and they require that the sums be barycentric. When a curve consists of a non-barycentric weighted sum of points, its shape depends on the particular coordinate system used. The shape changes when either the curve or the coordinate axes are moved or are affinely transformed. Such a curve is ill conditioned and cannot be used in practice.

Notice that the weights α and $(1-\alpha)$ in Equation (8.20) are barycentric. They are sometimes referred to as barycentric coordinates.

The term *barycentric* is derived from *barycenter*, meaning center of gravity, because such weights are used to calculate the center of gravity of an object. Barycentric weights have many applications in geometry in general and in curve and surface design in particular.

Note. The sum of points is generally undefined, but the special sum of two points $\mathbf{P}$ and $\mathbf{Q}$ can be written as the difference of points $\mathbf{P} + (-\mathbf{Q})$. Such a difference is the vector from $-\mathbf{Q}$ to $\mathbf{P}$ (Figure 8.41), which is why the sum $\mathbf{P} + (-\mathbf{Q})$ of two points is well defined. However, this sum tells us something about the relative positions of $\mathbf{P}$ and $-\mathbf{Q}$, not of $\mathbf{P}$ and $\mathbf{Q}$. Assuming that Figure 8.41 depicts the points $\mathbf{Q} = (-5, -1)$ and $\mathbf{P} = (4, 3)$, the sum $\mathbf{P} + \mathbf{Q}$ equals $(-5, -1) + (4, 3) = (-1, 2)$. This shows that in order to get from point $-\mathbf{Q}$ to point P, we need to move one negative step in the x direction for every two steps in the y direction.

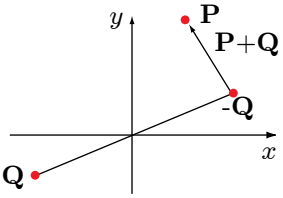


Figure 8.41: Adding Two Points.

Operations on vectors. Vectors can also be operated on, they can be added and subtracted, they can be multiplied by a real number, and there are also meaningful, well-defined ways to multiply two vectors. However, vector division cannot be defined in a way that makes sense or is useful or practical.

Adding or subtracting two vectors is done simply by adding or subtracting their individual components. Thus, for example, the sum or difference of $\mathbf{v} = (a, b)$ and $\mathbf{w} = (c, d)$ is $\mathbf{u} = (a \pm c, b \pm d)$, which is a vector. It does not have a location in space, but it has direction and magnitude like any other vector. The product $\alpha\mathbf{v}$ is a vector in the direction of $\mathbf{v}$ with magnitude $\alpha|\mathbf{v}|$.

The dot (inner or scalar) product of two vectors is denoted by $\mathbf{P} \bullet \mathbf{Q}$ and is defined as the scalar

$$(P_x, P_y, P_z)(Q_x, Q_y, Q_z)^T = \mathbf{P}\mathbf{Q}^T = P_xQ_x + P_yQ_y + P_zQ_z. \tag{8.22}$$

This also equals $|\mathbf{P}|\,|\mathbf{Q}|\cos\theta$, where θ is the angle between the vectors. The dot product of perpendicular vectors (also called orthogonal vectors) is therefore zero. The dot product is commutative, $\mathbf{P} \bullet \mathbf{Q} = \mathbf{Q} \bullet \mathbf{P}$.

The cross product (or vector product) of two vectors is denoted by $\mathbf{P} \times \mathbf{Q}$ and is defined as the vector

$$(P_2Q_3 - P_3Q_2, -P_1Q_3 + P_3Q_1, P_1Q_2 - P_2Q_1).$$

8.10 Representing Curves

There is an infinite variety of curves and they may differ significantly in their shapes and properties. Consider the differences between a straight line, which is smooth and proceeds in the same direction, a circle, which is fully curved and smooth, and a triangle, which features three sharp corners and straight segments. This is why mathematicians have developed three main approaches—explicit, implicit, and parametric—to mathematically represent curves

The explicit representation of curves is the familiar $y = f(x)$, which can be plotted as a curve. The advantage of this type of curve representation is its familiarity and ease of programming and plotting. However, the explicit representation is not general, because it is single-valued. For each value of x, only a single value of y is normally computed by the function. This makes it impossible to compute curves that have vertical parts, because all the points of a vertical segment have the same x coordinate and different y coordinates.

The implicit representation of a curve has the form $F(x, y) = 0$. It is general and can represent multivalued curves (more than one y value for an x value). A common example is the circle, whose implicit representation is $x^2+y^2-R^2 = 0$. On the downside, it may be hard to come up with the implicit expression of a given curve, and also difficult to write code for plotting it.

The explicit and implicit curve representations can be used only when the mathematical shape of the curve is known. In practical applications—where complex-looking curves such as the shape of a car or of a toaster are needed—the function is normally unknown, which is why a different approach is needed.

The curve representation used in practice is parametric. A two-dimensional parametric curve has the form $\mathbf{P}(t) = \big(f(t), g(t)\big)$ or $\mathbf{P}(t) = \big(x(t), y(t)\big)$. The functions f and g become the (x, y) coordinates of points on the curve, which can be computed and plotted when the parameter t is varied over a certain interval $[a, b]$, normally $[0, 1]$.

A basic example of a two-dimensional parametric curve is $\mathbf{P}(t) = (2t-1, t^2)$. When t is varied from 0 to 1, the curve proceeds from the initial point $\mathbf{P}(0) = (-1, 0)$ to the final point $\mathbf{P}(1) = (1, 1)$. The x coordinate is linear in t and the y coordinate varies as t^2.

The first derivative $\frac{d\mathbf{P}(t)}{dt}$ is denoted by $\mathbf{P}^t(t)$, or by $\dot{\mathbf{P}}$, or by $(P_x^t(t), P_y^t(t))$. This derivative is the tangent vector to the curve at any point. The derivative is a vector, not a number or a point, because it is the limit of the difference $(\mathbf{P}(t+\Delta) - \mathbf{P}(t))/\Delta$, and the difference of points is a vector. As a vector, the tangent possesses a direction (the direction of the curve at the point) and a magnitude (which indicates the speed of the curve at the point). The tangent, however, is not the slope of the curve. The tangent is a pair or a triplet of numbers, whereas the slope is a single number. The slope equals $\tan\theta$, where θ is the angle between the tangent vector and the x axis. The slope of a two-dimensional parametric curve is obtained by

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{P_y^t(t)}{P_x^t(t)}.$$

Example: The curve $\mathbf{P}(t) = (x(t), y(t)) = (1 + t^2/2, t^2)$. Its tangent vector is $\mathbf{P}^t(t) = (t, 2t)$ and its slope is $2t/t = 2$. The slope is constant, which indicates that the curve is a linear function, a straight line. This is also easy to see from the tangent vector. The direction of this vector is always the same, since it can be described by saying "for every t steps in the x direction, move $2t$ steps in the y direction."

Figure 8.42 illustrates examples of implicit and parametric curves, programmed in Geogebra.

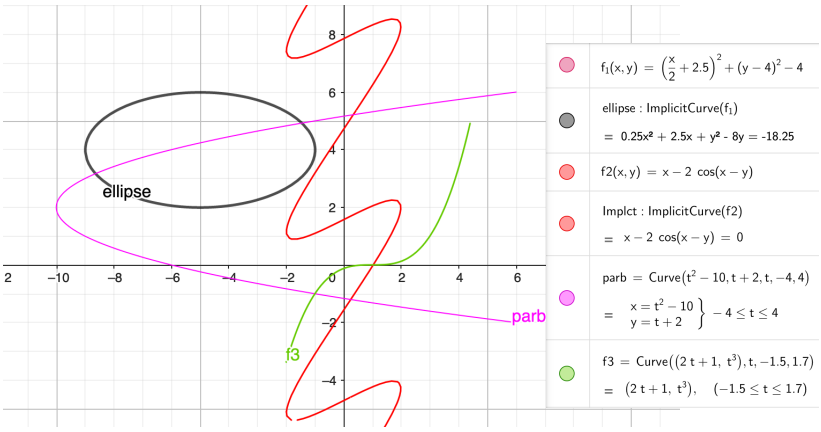


Figure 8.42: Implicit and Parametric Curves in Geogebra.

Note. Intuitively, the tangent line to a plane (two-dimensional) curve at a point $\mathbf{P}$ is the line that just "touches" the curve at $\mathbf{P}$. However, this definition is intuitive and applies only to plane curves. A rigorous, practical definition is needed that applies to both plane and space curves. We therefore say that the tangent vector to the parametric curve $\mathbf{P}(t)$ at point t_0 is the derivative with respect to t, $\dot{\mathbf{P}}(t_0)$, at point t_0. Notice that this derivative is a vector, not an infinite line, so it has a direction and magnitude which are functions of the same parameter t.

In the two-dimensional case, the curve also has a slope. If its tangent vector is $\mathbf{v} = (a, b)$, then the slope is b/a, and is undefined at points where a is zero (the tangent vector is vertical).

Tangent vectors and slopes are undefined where the curve is non-differentiable (has a sharp point).

Straight line—the shortest way between two points. —Euclid.
Cycloid—the fastest way between two points. —Johann Bernoulli.
Curve—the loveliest way between two points. —Mae West.

8.11 PC Curves

In late 20th century, with the proliferation of personal computers, the acronym PC stood for "personal computer." In this text, however, the term PC stands for parametric cubic. This is because many of the interpolation methods described here are based on parametric curves that are cubic polynomials. A polynomial of degree n in x is the function

$$P_n(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

where the a_i are the $n + 1$ (real) coefficients of the polynomial. PC curves are common in computer graphics applications because polynomials have the following features: (1) They are simple functions. (2) They are easy to compute, requiring only basic arithmetic operations. (3) They are defined everywhere, no singular points. (4) They are easy to differentiate and integrate. (5) Their coefficients appear linearly, i.e., in the form of a_i and not as a_i^2 or $\sqrt{a_i}$.

(A singularity is a point where the curve is undefined.)

Once we agree that parametric curves based on polynomials are useful for interpolation and curve design, we need to decide on the best degree for our polynomials. A polynomial of degree 1 has the form $\mathbf{P}_1(t) = \mathbf{A}t + \mathbf{B}$. It is a straight line, which is why its use is limited. A parametric polynomial of degree 2 (quadratic) has the form $\mathbf{P}_2(t) = \mathbf{A}t^2 + \mathbf{B}t + \mathbf{C}$, and the next paragraph shows that it is always a parabola. A parabola is not flexible enough for general use in practice because it can only have a limited range of shapes. In particular, it cannot have an inflection point (see below).

(A degree-2 parametric polynomial has one value for each value of the parameter t. It therefore cannot be a circle or an ellipse. It cannot be a hyperbola, because a hyperbola approaches ∞ and $-\infty$ in different directions, whereas a parabola approaches those two limits either going up or going down. It is true that plane curves described by degree-2 polynomials are conic sections, but this applies only to the implicit representation.)

◇ **Exercise 8.16:** Show that a quadratic polynomial must be a plane curve.

An inflection point is a point on a curve where the curvature is zero. On a straight line, every point is an inflection point. On a typical curve, an inflection point is created when the curve reverses its direction of turning (for example, from a clockwise direction to a counterclockwise direction). See Figure 8.43.

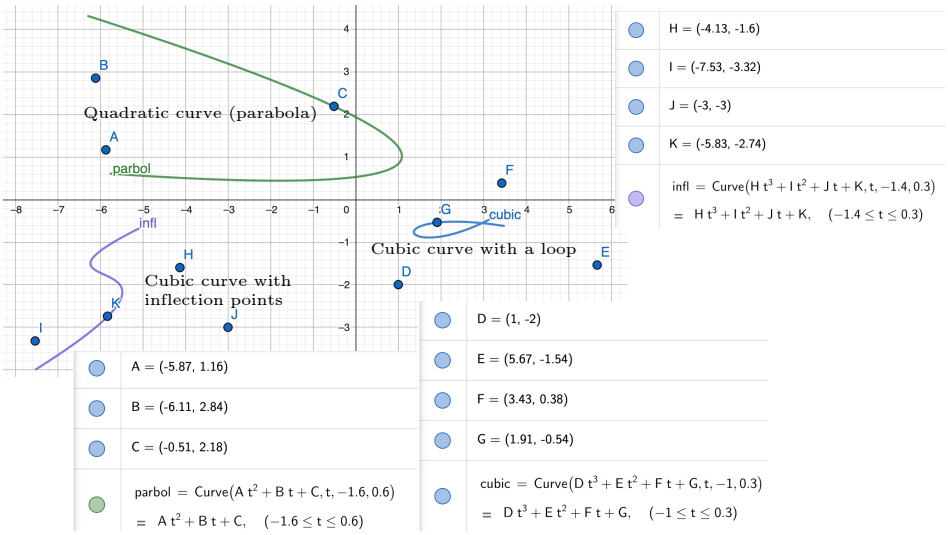
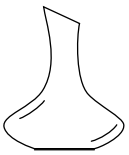


Figure 8.43: Three Parametric Curves.

A polynomial of degree 3 (cubic) has the form $\mathbf{P}_3(t) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D}$ and is the simplest curve that can have complex shapes and can also be a space curve. The complexity of this polynomial is limited, though. It can have at most one loop and, if it does not have a loop, it can have at most two inflection points. Polynomials of higher degrees are sometimes needed, but they generally wiggle too much, a feature known as Runge's phenomenon, and are difficult to control. They also have more coefficients, so they require more input data to determine the coefficients. As a result, a complete curve is often constructed from segments, each a parametric cubic polynomial (also called a PC). The complete curve is a piecewise polynomial curve that is termed a spline (Section 8.13).

◇ **Exercise 8.17:** Why does a high-degree polynomial wiggle?



Once we have decided to construct a long, complex curve from several short PC segments, each a cubic polynomial, we look at the mathematical details of such a segment and what is needed to compute it in practice. A practical way to compute the segments that combine to form the curves of Figure 8.44 is to start with data points, points through which the final curve would pass. Such points (red in the figure) are placed mostly at regions of high

curvature, and can be moved later, to edit and refine the shape of the curve. An alternative is to use control points (green in the figure). These are points that "pull" the curve toward them, and allow the designer to edit the curve in a predictable fashion. In addition to points, it is also possible to specify tangent vectors that determine the direction of the curve at certain points. The types of continuity of the curve at points where segments meet can also used to constrain the curve.

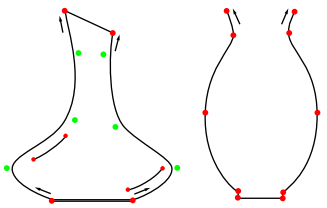


Figure 8.44: Data Points (Red), Control Points (Green), and Direction Vectors.

Regardless of the input data, a PC curve segment always has the form $\mathbf{P}(t) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D}$. Therefore, four unknown coefficients should be computed, which requires four equations. The equations must depend on four known quantities, points or vectors, that we denote by $\mathbf{G}_1$ through $\mathbf{G}_4$. The PC segment is expressed in compact matrix notation as the product

$$\mathbf{P}(t) = (t^3, t^2, t, 1) \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix} \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \mathbf{G}_3 \\ \mathbf{G}_4 \end{pmatrix} = \mathbf{T}(t) \cdot \mathbf{M} \cdot \mathbf{G},$$

where $\mathbf{M}$ is the basis matrix that depends on the method used and $\mathbf{G}$ is the geometry vector, which consists of the four given quantities. The segment can also be written as the weighted sum

$$\begin{aligned} \mathbf{P}(t) &= (t^3 m_{11} + t^2 m_{21} + t m_{31} + m_{41})\mathbf{G}_1 + (t^3 m_{12} + t^2 m_{22} + t m_{32} + m_{42})\mathbf{G}_2 \\ &\quad + (t^3 m_{13} + t^2 m_{23} + t m_{33} + m_{43})\mathbf{G}_3 + (t^3 m_{14} + t^2 m_{24} + t m_{34} + m_{44})\mathbf{G}_4 \\ &= B_1(t)\mathbf{G}_1 + B_2(t)\mathbf{G}_2 + B_3(t)\mathbf{G}_3 + B_4(t)\mathbf{G}_4 = \mathbf{B}(t) \cdot \mathbf{G} = \mathbf{T}(t) \cdot \mathbf{N} \cdot \mathbf{G}, \end{aligned}$$

where $\mathbf{B}(t)$ is the product $\mathbf{T}(t) \cdot \mathbf{M}$ and each $B_i(t)$ is the weight assigned to a $\mathbf{G}_i$. The weights are also referred to as blending functions, since they blend the four given quantities. If any of the quantities being blended are points, their weights should be barycentric. In the case where all four quantities are points, this requirement implies that the sum of the elements of matrix $\mathbf{M}$ should equal 1 (because the 16 elements of $\mathbf{M}$ are also the elements of the $B_i(t)$'s).

A PC segment can also be written in the form

$$\mathbf{P}(t) = \mathbf{A}t^3 + \mathbf{B}t^2 + \mathbf{C}t + \mathbf{D} = (t^3, t^2, t, 1) \begin{pmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \\ D_x & D_y & D_z \end{pmatrix} = \mathbf{T}(t) \cdot \mathbf{C},$$

where $\mathbf{A} = (A_x, A_y, A_z)$ and similarly for $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$. Its first derivative is

$$\frac{d\mathbf{P}(t)}{dt} = \frac{d\mathbf{T}(t)}{dt} \cdot \mathbf{C} = (3t^2, 2t, 1, 0)\mathbf{C}$$

and this is the tangent vector of the curve. This vector points in the direction of the tangent to the curve, but its magnitude is also important. It describes the speed of the curve.

Linear Interpolation (LERP)

Given a set of points, there are two ways to interpolate them linearly. If the plot of the points indicates that they are close to a straight line, we can employ the least-squares technique, Section 8.5, to compute the line that is the closest to all the points. Otherwise, we can calculate a set of straight segments (chords) connecting consecutive points. In the latter case, the order of the points is important.

Given two points $\mathbf{P}_1$ and $\mathbf{P}_2$, the straight segment connecting them is given by Equation (8.20), duplicated here

$$\mathbf{P}_1 = \mathbf{P}_0 + t(\mathbf{P}_2 - \mathbf{P}_0) = (1-t)\mathbf{P}_0 + t\mathbf{P}_2. \tag{8.20}$$

This equation can also be used to blend two arbitrary quantities, not just points in space. (A blend of two colors is a good example.) It creates a blend of t percent of $\mathbf{P}_2$ and $(1-t)$ percent of $\mathbf{P}_0$.

> A man who boasts about never changing his views is a man who's decided always to travel in a straight line—the kind of idiot who believes in absolutes.
>
> —Honoré de Balzac, *Père Goriot*, 1834.

8.12 Polynomial Interpolation

Given a set of $n+1$ ordered points, it is easy to compute a degree-n polynomial that passes through them, but the use of this interpolation method for curve design in practice is limited, because of the following reasons:

- As Exercise 8.17 shows, high-degree polynomials tend to wiggle and oscillate.
- The degree-n polynomial that passes through $n+1$ ordered points is unique. Such a polynomial is computed by a set of $n+1$ equations that are based on the known points. The equations yield $n+1$ solutions that become the coefficients of the polynomial, and the only way the user can interact with such a method is to move the points and recompute a new polynomial. If the points are given by a client or a boss, the curve designer may not be able to move them. If the resulting polynomial has the wrong shape, the designer must resort to other interpolation methods, such as splines, Section 8.13.

Two common methods for polynomial interpolation are the Lagrange and the Newton polynomials, and the former is discussed here.

The Lagrange Polynomial (LP)

This interpolation algorithm starts with a given ordered set of $n+1$ data points $\mathbf{P}_0 = (x_0, y_0)$, $\mathbf{P}_1 = (x_1, y_1), \ldots, \mathbf{P}_n = (x_n, y_n)$. For each point $\mathbf{P}_i$, an auxiliary polynomial $L_i^n(x)$ is defined that satisfies $L_i^n(x_i) = 1$, but is zero at any other data point. Once all the auxiliary polynomials are known, each $L_i^n(x)$ is multiplied by y_i and the sum $\sum y_i L_i^n(x)$ becomes the final result. It is obvious that there can be no editing by the user; the Lagrange polynomial is unique.

To construct a polynomial $P(x)$ that is zero at point a, b, and c, we simply include in it a product of the form $(x-a)(x-b)(x-c)$. Hence, the construction of $L_i^n(x)$ should start with the product

$$(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n). \tag{8.23}$$

In order for $L_i^n(x)$ to be 1 at x_i, we simply divide it by the value it otherwise would have at that point. From Equation (8.23), that value is

$$(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n).$$

The auxiliary polynomials are therefore

$$L_i^n(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} = \prod_{j=0, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)}.$$

Note that $(x - x_i)$ is missing from the numerator and $(x_i - x_i)$ is missing from the denominator. The final Lagrange polynomial (LP) is $\mathrm{LP}_n(x) = \sum_{i=0}^{n} y_i L_i^n(x)$.

◇ **Exercise 8.18:** Calculate the LP between the two points $\mathbf{P}_0 = (x_0, y_0)$ and $\mathbf{P}_1 = (x_1, y_1)$. What kind of a curve is it?

The LP can also be expressed in parametric form. Given the $n + 1$ data points $\mathbf{P}_0$, $\mathbf{P}_1, \ldots, \mathbf{P}_n$, we need to construct a polynomial $\mathbf{P}(t)$ that passes through all of them, such that $\mathbf{P}(t_0) = \mathbf{P}_0$, $\mathbf{P}(t_1) = \mathbf{P}_1$, …, $\mathbf{P}(t_n) = \mathbf{P}_n$, where $t_0 = 0$, $t_n = 1$, and t_1 through t_{n-1} are certain values between 0 and 1 (the t_i are called *knot* values). The LP has the form $\mathbf{P}(t) = \sum_{i=0}^{n} \mathbf{P}_i L_i^n(t)$. This is a weighted sum of the individual points where the weights (or basis functions) are given by

$$L_i^n(t) = \frac{\Pi_{j\neq i}^n (t - t_j)}{\Pi_{j\neq i}^n (t_i - t_j)}.$$

Note that $\sum_{i=0}^{n} L_i^n(t) = 1$, so these weights are barycentric. The LP is a rational polynomial because it is a quotient of two polynomials.

◇ **Exercise 8.19:** Calculate the parametric LP for the three points $\mathbf{P}_0 = (0,0)$, $\mathbf{P}_1 = (0,1)$, and $\mathbf{P}_2 = (1,1)$.

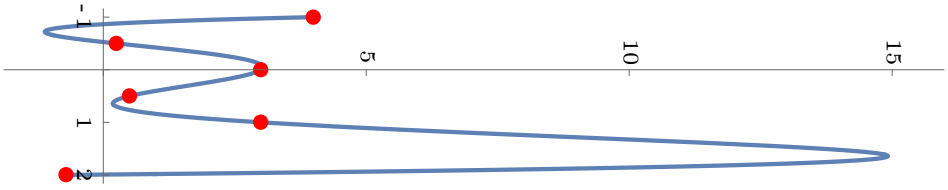
The quadratic Lagrange polynomial is especially useful and is often employed to interpolate three points. The auxiliary polynomials in this case are

$$\begin{aligned} L_0^2(t) &= \frac{\prod_{j\neq 0}^{2}(t-t_j)}{\prod_{j\neq 0}^{2}(t_0-t_j)} = \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)}, \\ L_1^2(t) &= \frac{\prod_{j\neq 1}^{2}(t-t_j)}{\prod_{j\neq 1}^{2}(t_1-t_j)} = \frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)}, \\ L_2^2(t) &= \frac{\prod_{j\neq 2}^{2}(t-t_j)}{\prod_{j\neq 2}^{2}(t_2-t_j)} = \frac{(t-t_0)(t-t_1)}{(t_2-t_0)(t_2-t_1)}, \end{aligned} \tag{8.24}$$

and the polynomial $\mathbf{P}_2(t) = \sum_{i=0}^{2} \mathbf{P}_i L_i^2(t)$ is easy to calculate and plot once the values of t_0, t_1, and t_2 have been determined. The choice $t_0 = 0$, $t_1 = 1$, and $t_2 = 2$ is particularly useful and yields the uniform quadratic Lagrange polynomial

$$\begin{aligned} \mathbf{P}_{2u}(t) &= \frac{t^2-3t+2}{2}\mathbf{P}_0 - (t^2-2t)\mathbf{P}_1 + \frac{t^2-t}{2}\mathbf{P}_2 \\ &= (t^2,t,1)\begin{pmatrix} 1/2 & -1 & 1/2 \\ -3/2 & 2 & -1/2 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{P}_0 \\ \mathbf{P}_1 \\ \mathbf{P}_2 \end{pmatrix}. \end{aligned}$$

Figure 8.45 illustrates a six-point Lagrange interpolation, lying on its side. The wild oscillations of the degree-5 polynomial are obvious.




```
pts={{-1,4},{2,-0.7},{0,3},{1,3},{0.5,0.5},{-0.5,0.25}};
InterpolatingPolynomial[pts,x]
g1=Plot[%, {x,-1,2}];
g2=ListPlot[pts,PlotStyle->{PointSize[Large],Red}];
Show[g1,g2,AspectRatio->Automatic,Ticks->{{-1,0,1,2},{0,5,10,15}},
 PlotRange->{{-1.1,2.1},{-1,15.1}}]

-0.7+(-2+x)(-1.56667+(1+x)(0.511111+(-0.5+x)(-1.86222+(-0.862222
 -7.76444(-1+x))(0.5 +x))))
```

Figure 8.45: A Six-Point Lagrange Interpolation Example.

The Newton Polynomial

The usefulness of the Lagrange polynomial is limited because it is unique. If the LP determined by the given data points is not what the designer wants, there is little that can be done to edit the interpolating curve and improve its form. The Newton interpolating polynomial is identical in its form to the Lagrange polynomial, but its derivation is different. It allows for adding more points without having to move the original data points. This offers the designer some control over the shape of the final curve.

We start with $n+1$ data points $\mathbf{P}_0$, $\mathbf{P}_1, \ldots, \mathbf{P}_n$, and assign them knot values

$$t_0 = 0 < t_1 < \cdots < t_{n-1} < t_n = 1.$$

From these, we are going to design a curve that is a degree-n parametric polynomial of the form

$$\mathbf{P}(t) = \sum_{i=0}^{n} N_i(t)\mathbf{A}_i,$$

where the basis functions $N_i(t)$ depend only on the knot values and not on the data points. Only the (unknown) coefficients $\mathbf{A}_i$ depend on the points. This definition, originally proposed by Newton, has the advantage that each coefficient $\mathbf{A}_i$ depends only on the "previous" points $\mathbf{P}_0$ through $\mathbf{P}_i$. If the user decides to add a point $\mathbf{P}_{i+1}$, only the coefficients and basis functions that follow, namely $\mathbf{A}_j$ and $N_j(t)$ for $j \geq i+1$, need be recomputed.

The definition of the basis functions is

$$N_0(t) = 1 \quad \text{and} \quad N_i(t) = (t-t_0)(t-t_1)\cdots(t-t_{i-1}), \quad \text{for} \quad i = 1, \ldots, n.$$

To calculate the unknown coefficients, we start with the sequence of equations

$$\begin{aligned}
\mathbf{P}_0 &= \mathbf{P}(t_0) = \mathbf{A}_0, \\
\mathbf{P}_1 &= \mathbf{P}(t_1) = \mathbf{A}_0 + \mathbf{A}_1(t_1 - t_0), \\
\mathbf{P}_2 &= \mathbf{P}(t_2) = \mathbf{A}_0 + \mathbf{A}_1(t_2 - t_0) + \mathbf{A}_2(t_2 - t_0)(t_2 - t_1), \\
&\vdots \\
\mathbf{P}_n &= \mathbf{P}(t_n) = \mathbf{A}_0 + \cdots.
\end{aligned}$$

These equations don't have to be solved simultaneously. Each can easily be solved after all its predecessors have been solved. The first few solutions are

$$\begin{aligned}
\mathbf{A}_0 &= \mathbf{P}_0, \\
\mathbf{A}_1 &= \frac{\mathbf{P}_1 - \mathbf{P}_0}{t_1 - t_0}, \\
\mathbf{A}_2 &= \frac{\mathbf{P}_2 - \mathbf{P}_0 - \dfrac{(\mathbf{P}_1 - \mathbf{P}_0)(t_2 - t_0)}{t_1 - t_0}}{(t_2 - t_0)(t_2 - t_1)} = \frac{\dfrac{\mathbf{P}_2 - \mathbf{P}_1}{t_2 - t_1} - \dfrac{\mathbf{P}_1 - \mathbf{P}_0}{t_1 - t_0}}{t_2 - t_0}.
\end{aligned}$$

This obviously becomes very complicated very quickly, which is why in practice, the method of divided differences [divided.diff 24] is used to express the solutions in compact notation. The divided difference of the knots $t_i t_k$ is denoted $[t_i t_k]$ and is defined as

$$[t_i t_k] \stackrel{\text{def}}{=} \frac{\mathbf{P}_i - \mathbf{P}_k}{t_i - t_k}.$$

The solutions can now be expressed as

$$\begin{aligned}
\mathbf{A}_0 &= \mathbf{P}_0, \\
\mathbf{A}_1 &= \frac{\mathbf{P}_1 - \mathbf{P}_0}{t_1 - t_0} = [t_1 t_0], \\
\mathbf{A}_2 &= [t_2 t_1 t_0] = \frac{[t_2 t_1] - [t_1 t_0]}{t_2 - t_0}, \\
\mathbf{A}_3 &= [t_3 t_2 t_1 t_0] = \frac{[t_3 t_2 t_1] - [t_2 t_1 t_0]}{t_3 - t_0}, \\
&\vdots \\
\mathbf{A}_n &= [t_n \dots t_1 t_0] = \frac{[t_n \dots t_1] - [t_{n-1} \dots t_0]}{t_n - t_0}.
\end{aligned}$$

◇ **Exercise 8.20:** Given the same points and knot values as in Exercise 8.19, calculate the Newton polynomial that passes through the points.

◇ **Exercise 8.21:** The tangent vector to a curve $\mathbf{P}(t)$ is the derivative $\frac{d\mathbf{P}(t)}{dt}$, which we denote by $\mathbf{P}^t(t)$. Calculate the tangent vectors to the curve of Exercises 8.19 and 8.20 at the three points. Also calculate the slopes of the curve at the points.

8.13 Spline Interpolation

Given a set of points, it is easy to compute a polynomial that passes through the points. The Lagrange polynomial (LP) on Page 478 is an example of such a polynomial. However, as the discussion in Section 8.11 (especially Exercise 8.17) illustrates, a curve based on a high-degree polynomial may wiggle wildly and its shape may be far from what the user has in mind. In practical work we are often interested in a smooth, tight curve that proceeds from point to point such that each segment between two points is a smooth arc. The spline approach to curve design, discussed in this chapter, constructs such a curve from individual segments, each a simple curve, generally a parametric cubic (PC).

Sections 8.17 through 8.19 illustrate spline interpolation techniques with three examples, cubic splines, cardinal splines and parabolic blending. One of the many available references for other types of splines is [Salomon 11]. A short history of splines plus much more information can be found in [WikiSpline 24].

Definition: A spline is a set of polynomials that are smoothly connected at certain data points. At each data point, two polynomials connect, and their first derivatives

(tangent vectors) have the same values. The definition also requires that all their derivatives up to the $(k-1)$st be the same at the point, thereby providing high-order continuity. There are exceptions to this definition. At some data points, the curve designer may want two segments to connect at a sharp corner. In some cases, only the first derivatives should have the same values (first-order continuity), or only the first and second derivatives should have the same values (second-order continuity), thereby resulting in a data point where two polynomial segments connect, but not very smoothly. Another point is that the polynomials that make up a spline may have different degrees. The degree of a spline is therefore defined as the maximum degree of its components.

Figure 8.46 is an example of three polynomials connecting four data points. The figure shows large parts of each polynomial, and it is clear that the spline requires only a small part of each. Thus, splines must have a way of using just a short segment of each of their polynomial components.

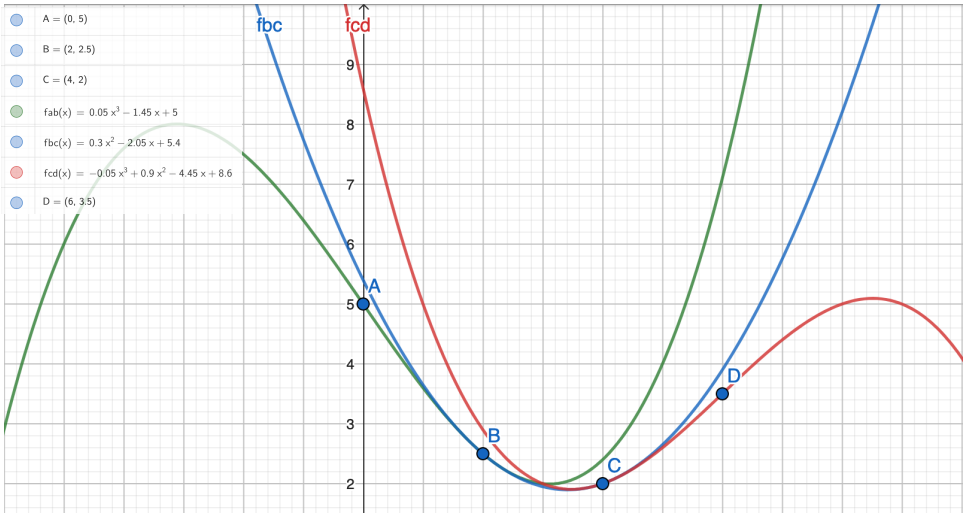


Figure 8.46: A Three-Part Cubic Spline.

As an introduction to spline techniques, the next section introduces the important interpolation method due to Charles Hermite, which serves as the basis for several other splines.

8.14 Hermite Interpolation

A practical curve design algorithm should be interactive. It should include user-controlled parameters that modify the shape of the curve in a predictable, intuitive way. The Hermite interpolation approach—developed by Charles Hermite (1822–1901) in the 1870's—is a good example of such a method.

> [Hermite] had a kind of positive hatred of geometry and once curiously reproached me with having made a geometrical memoir.
>
> —Jacques Hadamard.

Hermite interpolation is based on two points $\mathbf{P}_1$ and $\mathbf{P}_2$ and two direction vectors $\mathbf{P}_1^t$ and $\mathbf{P}_2^t$. It computes a curve segment that starts at $\mathbf{P}_1$ going in direction $\mathbf{P}_1^t$, and ends at $\mathbf{P}_2$ moving in direction $\mathbf{P}_2^t$. Hence, the direction vectors specified by the user become the tangent vectors at the ends of the Hermite curve segment. Several such curves are shown in Figure 8.47 with their endpoints and direction vectors.

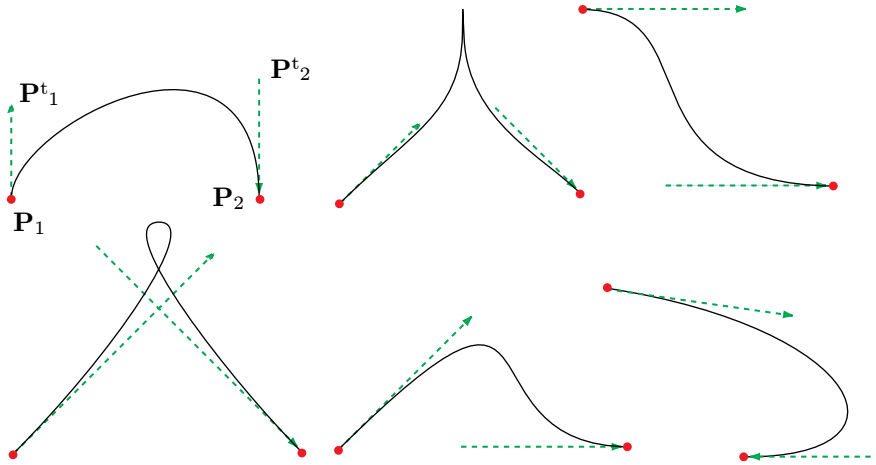


Figure 8.47: Various Hermite Curve Segments.

Figure 8.47 makes it clear that a Hermite segment can take on many different shapes, and can even have a cusp and develop a loop. A complete spline curve, however, normally requires several segments connected with different continuities. A C_0 continuity is the case where two segments are connected with a sharp corner. When two segments are connected with a C_1 continuity, their first derivatives (tangent vectors) are equal at the connection point. In general C_k continuity is the case where the first k derivatives are equal at the point.

8.15 Interactive Control

Hermite interpolation is interactive, which is an important advantage. If a Hermite curve segment does not look right, the user can edit it by modifying the direction vectors.

◇ **Exercise 8.22:** In the case of a four-point PC, we can change the shape of the curve by moving the points. Why then is the four-point method considered noninteractive?

Figure 8.47 illustrates how the shape of the curve depends on the directions of the direction vectors. Figure 8.48 shows how the curve can be edited by varying the magnitudes of those vectors. The figure shows three curves that start in a 45° direction and end up going vertically down. The effect illustrated here is simple. As the magnitude of the start direction vector increases, the curve continues longer in the original direction. This behavior implies that short direction vectors produce a curve that changes its direction early and aims straight toward the final point. Such a curve resembles a straight segment. The conclusion is that a long direction vector results in a loose curve and a short direction vector produces a tight curve.

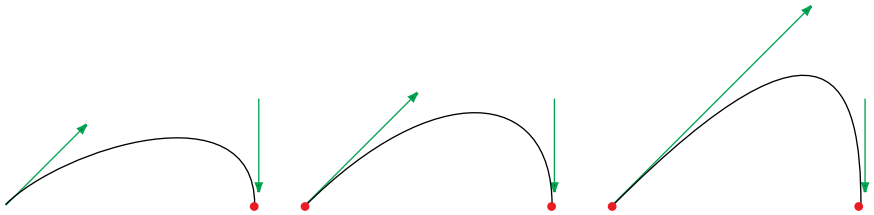


Figure 8.48: Varying the Direction Vector's Magnitude.

The reason the magnitudes, and not just the directions, of the vectors affect the shape of the curve is that the three-dimensional Hermite segment is a PC, and calculating a PC involves four coefficients, each a triplet, for a total of 12 unknown numbers. The two endpoints supply six known quantities and the two vectors should supply the remaining six. However, if we consider only the direction of a vector and not its magnitude, then the vectors $(1, 0.5, 0.3)$, $(2, 1, 0.6)$, and $(4, 2, 1.2)$ are identical. In such a case, only two of the three vector components are independent and two vectors provide only four independent quantities.

◇ **Exercise 8.23:** Discuss this claim in detail.

> A sketch tells as much in a glance as a dozen pages of print.
> —Ivan Turgenev, *Fathers and Sons* (1862).

8.16 The Hermite Curve Segment

The Hermite curve segment is easy to derive. It is a PC curve (a degree-3 parametric polynomial in t) with four coefficients that depend on the two points and two vectors. The basic equation of a PC curve is

$$\mathbf{P}(t) = \mathbf{a}t^3 + \mathbf{b}t^2 + \mathbf{c}t + \mathbf{d} = (t^3, t^2, t, 1)(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})^T = \mathbf{T}(t)\mathbf{A}. \tag{8.25}$$

This is the algebraic representation of the curve, in which the four coefficients are still unknown. Once these coefficients are expressed in terms of the known quantities, which are geometric, the curve will be expressed geometrically.

The tangent vector to a curve $\mathbf{P}(t)$ is the derivative $d\mathbf{P}(t)/dt$, which we denote by $\mathbf{P}^t(t)$. The tangent vector of a PC curve is therefore

$$\mathbf{P}^t(t) = 3\mathbf{a}t^2 + 2\mathbf{b}t + \mathbf{c}. \tag{8.26}$$

We denote the two given points by $\mathbf{P}_1$ and $\mathbf{P}_2$ and the two given tangents by $\mathbf{P}_1^t$ and $\mathbf{P}_2^t$. The four quantities are now used to calculate the geometric representation of the PC by writing equations that relate the four unknown coefficients $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$ to the four known ones, $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_1^t$, and $\mathbf{P}_2^t$. The equations are $\mathbf{P}(0) = \mathbf{P}_1$, $\mathbf{P}(1) = \mathbf{P}_2$, $\mathbf{P}^t(0) = \mathbf{P}_1^t$, and $\mathbf{P}^t(1) = \mathbf{P}_2^t$. Their explicit forms are

$$\begin{aligned} \mathbf{a}\cdot 0^3 + \mathbf{b}\cdot 0^2 + \mathbf{c}\cdot 0 + \mathbf{d} &= \mathbf{P}_1, \\ \mathbf{a}\cdot 1^3 + \mathbf{b}\cdot 1^2 + \mathbf{c}\cdot 1 + \mathbf{d} &= \mathbf{P}_2, \\ 3\mathbf{a}\cdot 0^2 + 2\mathbf{b}\cdot 0 + \mathbf{c} &= \mathbf{P}_1^t, \\ 3\mathbf{a}\cdot 1^2 + 2\mathbf{b}\cdot 1 + \mathbf{c} &= \mathbf{P}_2^t. \end{aligned} \tag{8.27}$$

They are easy to solve and the solutions are

$$\mathbf{a} = 2\mathbf{P}_1 - 2\mathbf{P}_2 + \mathbf{P}_1^t + \mathbf{P}_2^t, \quad \mathbf{b} = -3\mathbf{P}_1 + 3\mathbf{P}_2 - 2\mathbf{P}_1^t - \mathbf{P}_2^t, \quad \mathbf{c} = \mathbf{P}_1^t, \quad \mathbf{d} = \mathbf{P}_1. \tag{8.28}$$

Substituting these solutions into Equation (8.25) gives

$$\mathbf{P}(t) = (2\mathbf{P}_1 - 2\mathbf{P}_2 + \mathbf{P}_1^t + \mathbf{P}_2^t)t^3 + (-3\mathbf{P}_1 + 3\mathbf{P}_2 - 2\mathbf{P}_1^t - \mathbf{P}_2^t)t^2 + \mathbf{P}_1^t t + \mathbf{P}_1, \tag{8.29}$$

which, after rearranging, becomes

$$\begin{aligned} \mathbf{P}(t) &= (2t^3 - 3t^2 + 1)\mathbf{P}_1 + (-2t^3 + 3t^2)\mathbf{P}_2 + (t^3 - 2t^2 + t)\mathbf{P}_1^t + (t^3 - t^2)\mathbf{P}_2^t \\ &= F_1(t)\mathbf{P}_1 + F_2(t)\mathbf{P}_2 + F_3(t)\mathbf{P}_1^t + F_4(t)\mathbf{P}_2^t \\ &= (F_1(t), F_2(t), F_3(t), F_4(t))(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_1^t, \mathbf{P}_2^t)^T \\ &= \mathbf{F}(t)\mathbf{B}, \end{aligned} \tag{8.30}$$

where

$$\begin{aligned} F_1(t) = (2t^3 - 3t^2 + 1), \quad F_2(t) = (-2t^3 + 3t^2) = 1 - F_1(t), \\ F_3(t) = (t^3 - 2t^2 + t), \quad F_4(t) = (t^3 - t^2), \end{aligned} \tag{8.31}$$

$\mathbf{B}$ is the column $(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_1^t, \mathbf{P}_2^t)^T$, and $\mathbf{F}(t)$ is the row $(F_1(t), F_2(t), F_3(t), F_4(t))$. Equations (8.29) and (8.30) are the geometric representation of the Hermite PC segment.

Functions $F_i(t)$ are the Hermite blending functions. They create any point on the curve as a blend of the four given quantities. They are shown in Figure 8.49. Note that $F_1(t) + F_2(t) \equiv 1$. These two functions blend points, not tangent or direction vectors, and should therefore be barycentric. We can also write $F_1(t) = (t^3, t^2, t, 1)(2, -3, 0, 1)^T$ and similarly for $F_2(t)$, $F_3(t)$, and $F_4(t)$. In matrix notation this becomes

$$\mathbf{F}(t) = (t^3, t^2, t, 1) \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} = \mathbf{T}(t)\,\mathbf{H}.$$

The curve can now be written

$$\mathbf{P}(t) = \mathbf{F}(t)\mathbf{B} = \mathbf{T}(t)\,\mathbf{H}\,\mathbf{B} = (t^3, t^2, t, 1) \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_1^t \\ \mathbf{P}_2^t \end{pmatrix}. \tag{8.32}$$

Equation (8.25) tells us that $\mathbf{P}(t) = \mathbf{T}(t)\,\mathbf{A}$, which implies $\mathbf{A} = \mathbf{H}\,\mathbf{B}$. Matrix $\mathbf{H}$ is called the Hermite basis matrix.

The following is Mathematica code to display a single Hermite curve segment.

```
Clear[T,H,B]; (* Hermite Interpolation *)
T={t^3,t^2,t,1};
H={{2,-2,1,1},{-3,3,-2,-1},{0,0,1,0},{1,0,0,0}};
B={{0,0},{2,1},{1,1},{1,0}};
ParametricPlot[T.H.B,{t,0,1},PlotRange->All]
```

◇ **Exercise 8.24:** Express the midpoint $\mathbf{P}(0.5)$ of a Hermite segment in terms of the two endpoints and two tangent vectors. Draw a diagram to illustrate the geometric interpretation of the result.

8.16.1 Hermite Blending Functions

The four Hermite blending functions of Equation (8.31) are illustrated graphically in Figure 8.49. An analysis of these functions is essential for a thorough understanding of the Hermite interpolation method.

Function $F_1(t)$ is the weight assigned to the start point $\mathbf{P}_1$. It goes down from its maximum $F_1(0) = 1$ to $F_1(1) = 0$. This shows why for small values of t the curve is close to $\mathbf{P}_1$ and why $\mathbf{P}_1$ has little or no influence on the curve for large values of t. The opposite is true for $F_2(t)$, the weight of the endpoint $\mathbf{P}_2$. Function $F_3(t)$ is a bit trickier. It starts at zero, has a maximum at $t = 1/3$, then drops slowly back to zero. This behavior is interpreted as follows:

1. For small values of t, function $F_3(t)$ has almost no effect. The curve stays close to $\mathbf{P}_1$ regardless of the extreme tangents or anything else.

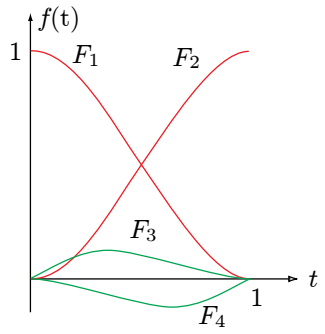


Figure 8.49: Hermite Weight Functions

2. For t values around 1/3, weight $F_3(t)$ exerts some influence on the curve. For these t values, weight $F_4(t)$ is small, and the curve is (approximately) the sum of (1) point $F_1(t)\mathbf{P}_1$ (large contribution), (2) point $F_2(t)\mathbf{P}_2$ (small contribution), and (3) vector $F_3(t)\mathbf{P}_1^t$. The sum of a point $\mathbf{P} = (x, y)$ and a vector $\mathbf{v} = (v_x, v_y)$ is a point located at $(x + v_x, y + v_y)$, which is how weight $F_3(t)$ "pulls" the curve in the direction of tangent (or direction) vector $\mathbf{P}_1^t$.

3. For large t values, function $F_3(t)$ again has almost no effect. The curve moves closer to $\mathbf{P}_2$ because weight $F_2(t)$ becomes dominant.

Function $F_4(t)$ is interpreted in a similar way. It has almost no effect for small and for large values of t. Its maximum (actually, minimum, because it is negative) occurs at $t = 2/3$, so it affects the curve only in this region. For t values close to 2/3, the curve is the sum of point $F_2(t)\mathbf{P}_2$ (large contribution), point $F_1(t)\mathbf{P}_1$ (small contribution), and vector $-|F_4(t)|\mathbf{P}_2^t$. Because $F_4(t)$ is negative, this sum is equivalent to $(x - v_x, y - v_y)$, which is why the curve approaches endpoint $\mathbf{P}_2$ while moving in direction $\mathbf{P}_2^t$.

Another important feature of the Hermite weight functions is that $F_1(t)$ and $F_2(t)$ are barycentric. They have to be, since they blend two points, and a detailed look at the four Equations (8.27) explains why they are. The first of these equations is simply $\mathbf{d} = \mathbf{P}_1$, which reduces the second one to $\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d} = \mathbf{P}_2$ or $\mathbf{a}+\mathbf{b}+\mathbf{c} = \mathbf{P}_2-\mathbf{P}_1$. The third equation solves $\mathbf{c}$, and the fourth equation, combined with the second equation, is finally used to compute $\mathbf{a}$ and $\mathbf{b}$. All this implies that $\mathbf{a}$ and $\mathbf{b}$ have the form $\mathbf{a} = \alpha(\mathbf{P}_2 - \mathbf{P}_1) + \cdots$, $\mathbf{b} = \beta(\mathbf{P}_2 - \mathbf{P}_1) + \cdots$. The final PC therefore has the form

$$\mathbf{P}(t) = \mathbf{a}t^3 + \mathbf{b}t^2 + \mathbf{c}t + \mathbf{d} = (\alpha\mathbf{P}_2 - \alpha\mathbf{P}_1 + \cdots)t^3 + (\beta\mathbf{P}_2 - \beta\mathbf{P}_1 + \cdots)t^2 + (\cdots)t + \mathbf{P}_1,$$

where the ellipses represent parts that depend only on the direction vectors, not on the endpoints. When this is rearranged, the result is

$$\mathbf{P}(t) = (-\alpha t^3 - \beta t^2 + 1)\mathbf{P}_1 + (\alpha t^3 + \beta t^2)\mathbf{P}_2 + (\cdots)\mathbf{P}_1^t + (\cdots)\mathbf{P}_2^t,$$

which is why the coefficients of $\mathbf{P}_1$ and $\mathbf{P}_2$ add up to unity.

8.16.2 Hermite Derivatives

The concept of blending can be applied to the calculation of the derivatives of a curve, not just to the curve itself. One way to calculate $\mathbf{P}^t(t)$ is to differentiate $\mathbf{T}(t) = (t^3, t^2, t, 1)$. The result is

$$\mathbf{P}^t(t) = \mathbf{T}^t(t)\mathbf{H}\mathbf{B} = (3t^2, 2t, 1, 0)\mathbf{H}\mathbf{B}.$$

A more general method is to use the relation $\mathbf{P}(t) = \mathbf{F}(t)\mathbf{B}$, which implies

$$\mathbf{P}^t(t) = \mathbf{F}^t(t)\mathbf{B} = \big(F_1^t(t), F_2^t(t), F_3^t(t), F_4^t(t)\big)\mathbf{B}.$$

The individual derivatives $F_i^t(t)$ can be obtained from Equation (8.31). The results can be expressed as

$$\mathbf{P}^t(t) = (t^3, t^2, t, 1)\begin{bmatrix} 0 & 0 & 0 & 0 \\ 6 & -6 & 3 & 3 \\ -6 & 6 & -4 & -2 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_1^t \\ \mathbf{P}_2^t \end{bmatrix} = \mathbf{T}(t)\mathbf{H}_t\mathbf{B}. \tag{8.33}$$

Similarly, the second derivatives of the Hermite segment can be expressed as

$$\mathbf{P}^{tt}(t) = (t^3, t^2, t, 1)\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 12 & -12 & 6 & 6 \\ -6 & 6 & -4 & -2 \end{bmatrix}\begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_1^t \\ \mathbf{P}_2^t \end{bmatrix} = \mathbf{T}(t)\mathbf{H}_{tt}\mathbf{B}. \tag{8.34}$$

These expressions make it easy to calculate the first and second derivatives at any point on a Hermite segment. Similar expressions can be derived for any other curves that are based on the blending of geometrical quantities.

◇ **Exercise 8.25:** What is H_{ttt}?

Example: The two two-dimensional points $\mathbf{P}_1 = (0, 0)$ and $\mathbf{P}_2 = (1, 0)$ and the two tangents $\mathbf{P}_1^t = (1, 1)$ and $\mathbf{P}_2^t = (0, -1)$ are given. The segment should therefore start at the origin, going in a 45° direction, and end at point $(1, 0)$, going straight down. The calculation of $\mathbf{P}(t)$ is straightforward:

$$\begin{aligned}
\mathbf{P}(t) &= \mathbf{T}(t)\,\mathbf{A} = \mathbf{T}(t)\,\mathbf{H}\,\mathbf{B} \\
&= (t^3, t^2, t, 1)\begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} (0,0) \\ (1,0) \\ (1,1) \\ (0,-1) \end{bmatrix} \\
&= (t^3, t^2, t, 1)\begin{bmatrix} 2(0,0) - 2(1,0) + 1(1,1) + 1(0,-1) \\ -3(0,0) + 3(1,0) - 2(1,1) - 1(0,-1) \\ 0(0,0) + 0(1,0) + 1(1,1) + 0(0,-1) \\ 1(0,0) + 0(1,0) + 0(1,1) + 0(0,-1) \end{bmatrix}
\end{aligned}$$

$$= (t^3, t^2, t, 1) \begin{bmatrix} (-1,0) \\ (1,-1) \\ (1,1) \\ (0,0) \end{bmatrix}$$
$$= (-1,0)t^3 + (1,-1)t^2 + (1,1)t. \tag{8.35}$$

◇ **Exercise 8.26:** Use Equation (8.35) to show that the segment really passes through points $(0,0)$ and $(1,0)$. Calculate the tangent vectors and use them to show that the segment really starts and ends in the right directions.

◇ **Exercise 8.27:** Repeat the example above with $\mathbf{P}_1^t = (2,2)$. The new curve segment should go through the same points, in the same directions. However, it should continue longer in the original 45° direction, since the size of the new tangent is $\sqrt{2^2+2^2} = 2\sqrt{2}$, twice as long as the previous one, which is $\sqrt{1^2+1^2} = \sqrt{2}$.

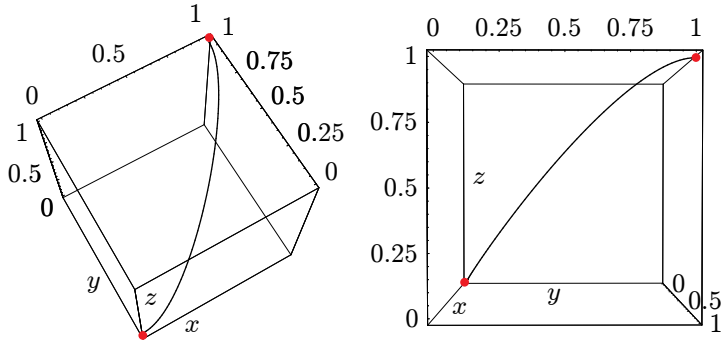
◇ **Exercise 8.28:** Calculate the Hermite curve for two given points $\mathbf{P}_1$ and $\mathbf{P}_2$ assuming that the tangent vectors at the two points are zero (indeterminate). What kind of a curve is this?

◇ **Exercise 8.29:** Use the Hermite method to calculate PC segments for the cases where the known quantities are as follows:

1. The three tangent vectors at the start, middle, and end of the segment.
2. The two interior points $\mathbf{P}(1/3)$ and $\mathbf{P}(2/3)$, and the two extreme tangent vectors $\mathbf{P}^t(0)$ and $\mathbf{P}^t(1)$.
3. The two extreme points $\mathbf{P}(0)$ and $\mathbf{P}(1)$, and the two interior tangent vectors $\mathbf{P}^t(1/3)$ and $\mathbf{P}^t(2/3)$ (this is similar to case 2, so it's easy).

Example: Given the two three-dimensional points $\mathbf{P}_1 = (0,0,0)$ and $\mathbf{P}_2 = (1,1,1)$, and the two tangent vectors $\mathbf{P}_1^t = (1,0,0)$ and $\mathbf{P}_2^t = (0,1,0)$, the curve segment is the simple cubic polynomial shown in Figure 8.50

$$\mathbf{P}(t) = (t^3, t^2, t, 1) \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} (0,0,0) \\ (1,1,1) \\ (1,0,0) \\ (0,1,0) \end{bmatrix}$$
$$= (-t^3 + t^2 + t, -t^3 + 2t^2, -2t^3 + 3t^2). \tag{8.36}$$




```
(* Hermite 3D example *)
Clear[T,H,B];
T={t^3,t^2,t,1};
H={{2,-2,1,1},{-3,3,-2,-1},{0,0,1,0},{1,0,0,0}};
B={{0,0,0},{1,1,1},{1,0,0},{0,1,0}};
ParametricPlot3D[T.H.B,{t,0,1},
 ViewPoint->{-0.846, -1.464, 3.997}];
(* ViewPoint->{3.119, -0.019, 0.054} alt view *)
```

Figure 8.50: A Hermite Curve Segment in Space.

8.17 The Cubic Spline Curve

The cubic spline was originally introduced by James Ferguson in 1964. Given the n data points $\mathbf{P}_1$ through $\mathbf{P}_n$, this method computes a curve that passes through the points and whose shape is often pleasing to the eye. There are infinitely many curves that pass through any set of ordered points in order of their numbers (Figure 8.51a), but the eye often tends to trace *one* imaginary smooth curve through the points, especially if the points are arranged in a familiar pattern. It is therefore useful to have an algorithm that does the same and is also interactive, thereby giving the user a chance to modify the curve by making changes, but without moving points.

The cubic spline curve consists of $n-1$ individual Hermite segments that are smoothly connected at the $n-2$ interior points and that are easy to compute and display. In order for the segments to meet at the interior points, their tangent vectors (first derivatives) must be the same at each interior point. An added feature of cubic splines is that their second derivatives are also equal at the interior points. The cubic spline method is interactive. The user can control the curve by varying the tangent vectors at the two ends of the curve.

Given the n data points $\mathbf{P}_1$ through $\mathbf{P}_n$, we look for $n-1$ parametric cubics $\mathbf{P}_1(t)$, $\mathbf{P}_2(t)$, ..., $\mathbf{P}_{n-1}(t)$, such that $\mathbf{P}_k(t)$ is the polynomial segment from point $\mathbf{P}_k$ to point $\mathbf{P}_{k+1}$ (Figure 8.51b). The PCs would have to be smoothly connected at the $n-2$ interior points $\mathbf{P}_2$ through $\mathbf{P}_{n-1}$, which means that their first derivatives would have to match at every interior point. The definition of a spline demands that their second derivatives match too. This requirement (the boundary condition of the cubic spline) provides the

necessary equations and also results in a tight curve, in the sense that once the entire curve is drawn, the eye can no longer detect the positions of the original data points.

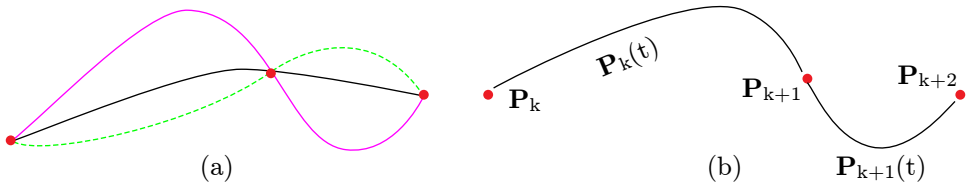


Figure 8.51: (a) Three Different Curves. (b) Two Segments.

The principle of cubic splines is to divide the set of n points into $n-1$ overlapping pairs of two points each and to fit a Hermite segment (Equations (8.29) and (8.30)) to each pair. The pairs are $(\mathbf{P}_1, \mathbf{P}_2)$, $(\mathbf{P}_2, \mathbf{P}_3)$, up to $(\mathbf{P}_{n-1}, \mathbf{P}_n)$. Recall that a Hermite curve segment is specified by two points and two tangents. In our case, all the points are given, by a client or a boss. They are known and cannot be moved by the curve designer. The only unknowns are the tangent vectors. In order for segments $\mathbf{P}_k(t)$ and $\mathbf{P}_{k+1}(t)$ to connect smoothly at point $\mathbf{P}_{k+1}$, the end tangent of $\mathbf{P}_k(t)$ must equal the start tangent of $\mathbf{P}_{k+1}(t)$. Thus, there is only one tangent vector per point, for a total of n unknowns.

The unknown tangent vectors are computed as the solutions of a system of n equations. The equations are derived from the requirement that the second derivatives of the individual segments match at every interior point. However, there are only $n-2$ interior points, so we can only have $n-2$ equations, enough to solve for only $n-2$ unknowns.

The key to resolving this shortage of equations is to ask the user to provide the software with the values of two tangent vectors, normally the ones at the extreme ends of the curve. Once this is done, the equations are solved, yielding the remaining $n-2$ tangents. This seems a strange way to solve equations, but it has the advantage of being interactive. If the resulting curve looks wrong, the user can repeat the calculation with two new extreme tangent vectors. Before delving into the details, here is a summary of the steps involved.

1. The n data points are input into the software.
2. The user provides values (guesses or estimates) for two tangent vectors.
3. The software sets up $n-2$ equations, with the remaining $n-2$ tangent vectors as the unknowns, and solves them.
4. The software loops $n-1$ times. In each iteration, it selects two adjacent points and their tangent vectors to compute one Hermite segment.

We start with three adjacent points, $\mathbf{P}_k$, $\mathbf{P}_{k+1}$, and $\mathbf{P}_{k+2}$, of which $\mathbf{P}_{k+1}$ is an interior point and the other two can be either interior or endpoints. Thus, k varies from 1 to $n-2$. The Hermite segment from $\mathbf{P}_k$ to $\mathbf{P}_{k+1}$ is denoted by $\mathbf{P}_k(t)$, which implies that $\mathbf{P}_k(0) = \mathbf{P}_k$ and $\mathbf{P}_k(1) = \mathbf{P}_{k+1}$. The tangent vectors of $\mathbf{P}_k(t)$ at the endpoints are still unknown and will be denoted by $\mathbf{P}_k^t$ and $\mathbf{P}_{k+1}^t$. The first step is to express segment $\mathbf{P}_k(t)$ geometrically, in terms of the two endpoints and the two tangents.

Applying Equation (8.29) to our segment results in

$$\begin{aligned}\mathbf{P}_k(t) = \mathbf{P}_k + \mathbf{P}_k^t t + \left[3(\mathbf{P}_{k+1} - \mathbf{P}_k) - 2\mathbf{P}_k^t - \mathbf{P}_{k+1}^t\right] t^2 \\ + \left[2(\mathbf{P}_k - \mathbf{P}_{k+1}) + \mathbf{P}_k^t + \mathbf{P}_{k+1}^t\right] t^3.\end{aligned} \tag{8.37}$$

When the same equation is applied to the next segment $\mathbf{P}_{k+1}(t)$ (from $\mathbf{P}_{k+1}$ to $\mathbf{P}_{k+2}$), it becomes

$$\begin{aligned}\mathbf{P}_{k+1}(t) = \mathbf{P}_{k+1} + \mathbf{P}_{k+1}^t t + \left[3(\mathbf{P}_{k+2} - \mathbf{P}_{k+1}) - 2\mathbf{P}_{k+1}^t - \mathbf{P}_{k+2}^t\right] t^2 \\ + \left[2(\mathbf{P}_{k+1} - \mathbf{P}_{k+2}) + \mathbf{P}_{k+1}^t + \mathbf{P}_{k+2}^t\right] t^3.\end{aligned} \tag{8.38}$$

◇ **Exercise 8.30:** Where do we use the assumption that the first derivatives of segments $\mathbf{P}_k(t)$ and $\mathbf{P}_{k+1}(t)$ are equal at the interior point $\mathbf{P}_{k+1}$?

Next, we use the requirement that the second derivatives of the two segments be equal at the interior points. The second derivative $\mathbf{P}^{tt}(t)$ of a Hermite segment $\mathbf{P}(t)$ is obtained by differentiating Equation (8.26)

$$\mathbf{P}^{tt}(t) = 6\mathbf{a}t + 2\mathbf{b}. \tag{8.39}$$

Equality of the second derivatives at the interior point $\mathbf{P}_{k+1}$ implies

$$\mathbf{P}_k^{tt}(1) = \mathbf{P}_{k+1}^{tt}(0) \qquad \text{or} \qquad 6\mathbf{a}_k \times 1 + 2\mathbf{b}_k = 6\mathbf{a}_{k+1} \times 0 + 2\mathbf{b}_{k+1}. \tag{8.40}$$

Using the values of $\mathbf{a}$ and $\mathbf{b}$ from Equations (8.37) and (8.38), we get

$$\begin{aligned}6\left[2(\mathbf{P}_k - \mathbf{P}_{k+1}) + \mathbf{P}_k^t + \mathbf{P}_{k+1}^t\right] + 2\left[3(\mathbf{P}_{k+1} - \mathbf{P}_k) - 2\mathbf{P}_k^t - \mathbf{P}_{k+1}^t\right] \\ = 2\left[3(\mathbf{P}_{k+2} - \mathbf{P}_{k+1}) - 2\mathbf{P}_{k+1}^t - \mathbf{P}_{k+2}^t\right],\end{aligned} \tag{8.41}$$

which, after simple algebraic manipulations, becomes

$$\mathbf{P}_k^t + 4\mathbf{P}_{k+1}^t + \mathbf{P}_{k+2}^t = 3(\mathbf{P}_{k+2} - \mathbf{P}_k). \tag{8.42}$$

The three quantities on the left side of Equation (8.42) are unknown. The two quantities on the right side are known.

Equation (8.42) can be written $n-2$ times for all the interior points $\mathbf{P}_{k+1} = \mathbf{P}_2, \mathbf{P}_3, \ldots, \mathbf{P}_{n-1}$ to obtain a system of $n-2$ linear algebraic equations expressed in matrix form as

$$\underbrace{\begin{pmatrix} 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 \\ & & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & 4 & 1 \end{pmatrix}}_{n} \begin{pmatrix} \mathbf{P}_1^t \\ \mathbf{P}_2^t \\ \vdots \\ \mathbf{P}_n^t \end{pmatrix} = \begin{pmatrix} 3(\mathbf{P}_3 - \mathbf{P}_1) \\ 3(\mathbf{P}_4 - \mathbf{P}_2) \\ \vdots \\ 3(\mathbf{P}_n - \mathbf{P}_{n-2}) \end{pmatrix}. \tag{8.43}$$

$n-2$

Equation (8.43) is a system of $n-2$ equations in the n quantities $\mathbf{P}_1^t$ through $\mathbf{P}_n^t$. We assume that two of these quantities namely, the two extreme tangents $\mathbf{P}_1^t$ and $\mathbf{P}_n^t$, have already been input by the user, leaving only $n-2$ unknowns. Equation (8.43) is now easily solved, yielding the values of the remaining $n-2$ tangents, $\mathbf{P}_2^t$ through $\mathbf{P}_{n-1}^t$. The n tangent vectors are now used to calculate the original coefficients **a**, **b**, **c**, and **d** of each segment by means of Equations (8.28), (8.29), or (8.32), which should be written and solved $n-1$ times, once for each segment of the spline.

The reader should notice that the coefficient matrix of Equation (8.43) is tridiagonal and therefore diagonally dominant and thus nonsingular. This guarantees that our system of equations can always be solved and that it has a unique solution.

This approach to solving Equation (8.43) is called the *clamped* end condition. Its advantage is that the user can vary the shape of the curve by entering new values for $\mathbf{P}_1^t$ and $\mathbf{P}_n^t$ and recalculating. This allows for interactive design, where each step brings the curve closer to the desired shape. Figure 8.51a is an example of three cubic splines that pass through the same points and differ only in $\mathbf{P}_1^t$ and $\mathbf{P}_n^t$. It illustrates how the shape of the entire curve can be radically modified by varying the two extreme tangents.

It is possible to let the user specify any two tangent vectors, not just the two extreme ones. However, varying the two extreme tangents is a natural way to edit and reshape the curve in practical applications.

Tension control. A tension parameter can be introduced in order to control the tension of a cubic spline. This is done by varying the magnitudes of the tangent vectors of the individual Hermite segments that constitute the spline. The user inputs a tension parameter s and the software multiplies every tangent vector by s. Small values of s correspond to short vectors and thus to high tension (a tight curve, close to a straight line). Large s results in a loose, low-tension spline. A user-friendly algorithm should define a parameter T in the interval $[0,1]$ and should multiply each tangent vector by $s=\alpha(1-T)$ for some predetermined α. Large values of T (close to 1) correspond to small s and therefore to high tension, while small values of T correspond to s close to α. This makes T a natural tension parameter. Section 8.18 introduces the similar relation $T=1-2s$, which makes more sense for cardinal splines.

The downside of the cubic spline is the following:

1. There is no local control. Modifying the extreme tangent vectors changes Equation (8.43) and results in a different set of n tangent vectors. The entire curve is modified!

2. Equation (8.43) is a system of n equations that, for large values of n, may be too slow to solve.

Picnic Blues (anagram of Cubic Spline).

8.17.1 Example

Given the four points $\mathbf{P}_1=(0,0)$, $\mathbf{P}_2=(1,0)$, $\mathbf{P}_3=(2,0)$, and $\mathbf{P}_4=(3,0)$, we are looking for three Hermite segments $\mathbf{P}_1(t)$, $\mathbf{P}_2(t)$, and $\mathbf{P}_3(t)$ that will connect smoothly at the two interior points $\mathbf{P}_2$ and $\mathbf{P}_3$ and will constitute the spline. We further select an initial direction $\mathbf{P}_1^t=(1,-1)$ and a final direction $\mathbf{P}_4^t=(-1,-1)$.

We first write Equation (8.43) for our special case ($n = 4$)

$$\begin{pmatrix} 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} (1,-1) \\ \mathbf{P}_2^t \\ \mathbf{P}_3^t \\ (-1,-1) \end{pmatrix} = \begin{pmatrix} 3[(2,0)-(0,0)] \\ 3[(3,0)-(1,0)] \end{pmatrix}.$$

This is a system of two equations in two unknowns. It is easy to solve and the solutions are $\mathbf{P}_2^t = (\frac{13}{15}, \frac{1}{5})$ and $\mathbf{P}_3^t = (\frac{23}{15}, \frac{1}{5})$.

We now write Equation (8.32) three times, for the three spline segments. For the first segment, Equation (8.32) becomes

$$\mathbf{P}_1(t) = (t^3, t^2, t, 1) \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} (0,0) \\ (1,0) \\ (1,-1) \\ (\frac{13}{15}, \frac{1}{5}) \end{pmatrix}.$$

The second segment is calculated in a similar way:

$$\mathbf{P}_2(t) = (t^3, t^2, t, 1) \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} (1,0) \\ (2,0) \\ (\frac{13}{15}, \frac{1}{5}) \\ (\frac{23}{15}, \frac{1}{5}) \end{pmatrix}.$$

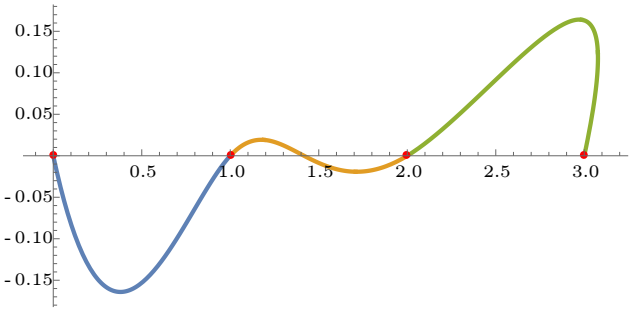
Finally, we write, for the third segment,

$$\mathbf{P}_3(t) = (t^3, t^2, t, 1) \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} (2,0) \\ (3,0) \\ (\frac{23}{15}, \frac{1}{5}) \\ (-1,-1) \end{pmatrix},$$

which completes the example. Figure 8.52 shows the points, the two extreme tangent vectors, and the resulting curve.

◇ **Exercise 8.31:** Redo the example of this section with an indefinite initial direction $\mathbf{P}_1^t = (0,0)$. What does it mean for a curve to start going in an indefinite direction?

Reference [Salomon 11] has much more information on the following special types of cubic splines: Relaxed, cyclic, periodic, closed, nonuniform, and fair.




```
ClearAll[p1,p2,p3,p4,pt1,pt2,pt3,pt4];
p1={0,0}; p2={1,0}; p3={2,0}; p4={3,0};
mt={{1,4,1,0},{0,1,4,1}};
c={{{{1,-1}}},{{{a11,a12}}},{{{a21,a22}}},{{{-1,-1}}}};
r={{3({p3}-{p1})},{3({p4}-{p2})}};
Solve[mt.c==r, {a11,a12,a21,a22}]
pt1={1,-1};pt2={13/15,1/5};
pt3={23/15,1/5};pt4={-1,-1};
tr={t^3,t^2,t,1};
m4={{2,-2,1,1},{-3,3,-2,-1},{0,0,1,0},{1,0,0,0}};
po1[t_]:=tr.m4.{{{p1}},{{p2}},{{pt1}},{{pt2}}};
po2[t_]:=tr.m4.{{{p2}},{{p3}},{{pt2}},{{pt3}}};
po3[t_]:=tr.m4.{{{p3}},{{p4}},{{pt3}},{{pt4}}};
g1=ParametricPlot[{po1[t],po2[t],po3[t]},{t,0,1}];
g2=ListPlot[{p1,p2,p3,p4},PlotStyle->{PointSize[Medium],Red}];
Show[g1,g2,AspectRatio->.5]

{{a11->13/15,a12->1/5,a21->23/15,a22->1/5}}
```

Figure 8.52: A Cubic Spline Example.

8.18 Cardinal Splines

The cardinal spline is an example of how Hermite interpolation can be applied to construct a spline. The cardinal spline offers local control as well as a natural way to control the tension of the curve by modifying the magnitudes of certain direction vectors. The price for all this is the loss of second-order continuity. At a data point where two polynomial segments meet, their first derivatives are equal, but their second derivatives are not.

Figure 8.53a illustrates the principle of this method. The figure shows seven points and a seemingly continuous curve that consists of four segments and depends on certain (green dashed) vectors. There are no segments connecting the first and last points. In general, a cardinal spline starts with n points. It uses $n-2$ direction vectors, each pointing from a data point $\mathbf{P}_i$ to point $\mathbf{P}_{i+2}$. The spline also consists of $n-3$ polynomial segments. The first segment, from $\mathbf{P}_2$ to $\mathbf{P}_3$, starts in the direction from $\mathbf{P}_1$ to $\mathbf{P}_3$ and ends going in the direction from $\mathbf{P}_2$ to $\mathbf{P}_4$. The last segment, from $\mathbf{P}_5$ to $\mathbf{P}_6$, starts in the direction from $\mathbf{P}_4$ to $\mathbf{P}_6$ and ends going in the direction from $\mathbf{P}_5$ to $\mathbf{P}_7$.

Each polynomial segment of a cardinal spline depends on four points, which is why there are only $n-3$ segments for the n data points. Similarly, each data point

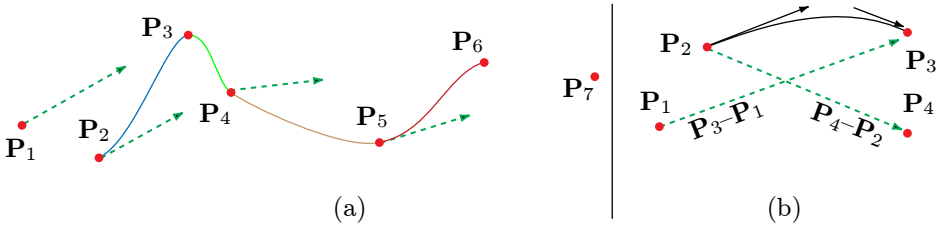


Figure 8.53: Direction Vectors in a Cardinal Spline.

participates in at most four curve segments, so moving one point affects only those segments and not the entire spline. This is why the cardinal spline method features *local control*. The individual segments connect smoothly, and their first derivatives are equal at the connection points (the curve features first-order continuity). However, the second derivatives of the segments are generally different at the connection points.

The first step in constructing the complete curve is to organize the points into $n-3$ highly-overlapping groups of four consecutive points each. The groups are

$$[\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3,\mathbf{P}_4],\quad [\mathbf{P}_2,\mathbf{P}_3,\mathbf{P}_4,\mathbf{P}_5],\quad [\mathbf{P}_3,\mathbf{P}_4,\mathbf{P}_5,\mathbf{P}_6],\ldots,[\mathbf{P}_{n-3},\mathbf{P}_{n-2},\mathbf{P}_{n-1},\mathbf{P}_n].$$

Hermite interpolation is then applied to construct a curve segment $\mathbf{P}(t)$ for each group. Denoting the four points of a group by $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$, and $\mathbf{P}_4$, the two interior points $\mathbf{P}_2$ and $\mathbf{P}_3$ become the start and end points of the segment and the two direction vectors become $s(\mathbf{P}_3-\mathbf{P}_1)$ and $s(\mathbf{P}_4-\mathbf{P}_2)$, where s, discussed below, is a real number related to the tension of the curve. Thus, segment $\mathbf{P}(t)$ goes from $\mathbf{P}_2$ to $\mathbf{P}_3$ and its two extreme direction vectors are proportional to the vectors $\mathbf{P}_3-\mathbf{P}_1$ and $\mathbf{P}_4-\mathbf{P}_2$ (Figure 8.53b). The proportionality constant s is related to the tension parameter T. Note how there are no segments from $\mathbf{P}_1$ to $\mathbf{P}_2$ and from $\mathbf{P}_{n-1}$ to $\mathbf{P}_n$. The two extreme points are used to edit the curve, because the first segment, from $\mathbf{P}_1$ to $\mathbf{P}_2$, starts going in the direction from $\mathbf{P}_0$ to $\mathbf{P}_2$, and similarly for the last segment.

The particular choice of the direction vectors guarantees that the individual segments of the cardinal spline connect smoothly. The end direction $s(\mathbf{P}_4-\mathbf{P}_2)$ of the segment for group $[\mathbf{P}_1,\mathbf{P}_2,\mathbf{P}_3,\mathbf{P}_4]$ is identical to the start direction of the next group, $[\mathbf{P}_2,\mathbf{P}_3,\mathbf{P}_4,\mathbf{P}_5]$.

Segment $\mathbf{P}(t)$ is therefore defined by

$$\begin{aligned}\mathbf{P}(0)=\mathbf{P}_2,&\quad \mathbf{P}(1)=\mathbf{P}_3,\\ \mathbf{P}^t(0)=s(\mathbf{P}_3-\mathbf{P}_1),&\quad \mathbf{P}^t(1)=s(\mathbf{P}_4-\mathbf{P}_2)\end{aligned}\tag{8.44}$$

and is easily calculated by applying Hermite interpolation (Equation (8.32)) to the four quantities of Equation (8.44)

$$\mathbf{P}(t)=(t^3,t^2,t,1)\begin{pmatrix}2&-2&1&1\\-3&3&-2&-1\\0&0&1&0\\1&0&0&0\end{pmatrix}\begin{pmatrix}\mathbf{P}_2\\\mathbf{P}_3\\s(\mathbf{P}_3-\mathbf{P}_1)\\s(\mathbf{P}_4-\mathbf{P}_2)\end{pmatrix}$$

$$= (t^3, t^2, t, 1) \begin{pmatrix} -s & 2-s & s-2 & s \\ 2s & s-3 & 3-2s & -s \\ -s & 0 & s & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_3 \\ \mathbf{P}_4 \end{pmatrix}. \tag{8.45}$$

Tension in the cardinal spline can now be controlled by varying the lengths of the direction vectors by means of parameter s. A long direction vector (obtained by a large s) causes the curve to continue longer in the direction of the vector. A short direction vector has the opposite effect; the curve moves a short distance in the direction of the vector, then quickly changes direction and moves toward the end point. A zero-length vector (corresponding to $s = 0$) produces a straight line between the endpoints (infinite tension). In principle, the parameter s can be varied from 0 to ∞. In practice, we use only values in the range $[0, 1]$. However, since $s = 0$ produces maximum tension, we cannot intuitively think of s as the tension parameter and another parameter, T, needs to be defined, which is inversely related to s.

The tension parameter T is defined as $s = (1 - T)/2$, which implies $T = 1 - 2s$. The value $T = 0$ results in $s = 1/2$. The curve is defined as having tension zero in this case and is called the Catmull-Rom spline [Catmull and Rom 74]. Section 8.19 includes a detailed derivation of this type of spline as a blend of two parabolas. Increasing T from 0 to 1 decreases s from 1/2 to 0, thereby reducing the magnitude of the direction vectors down to 0. This produces curves with more tension. We know that when the direction vectors have magnitude zero, the Hermite curve segment is a straight line, so the entire cardinal spline curve becomes a set of straight segments, a polyline, the curve with maximum tension. Decreasing T from 0 to -1 increases s from 1/2 to 1. The result is a curve with more slack at the data points.

To illustrate this behavior mathematically, we rewrite Equation (8.45) explicitly to show its dependence on s:

$$\begin{aligned} \mathbf{P}(t) = {} & s(-t^3 + 2t^2 - t)\mathbf{P}_1 + s(-t^3 + t^2)\mathbf{P}_2 + (2t^3 - 3t^2 + 1)\mathbf{P}_2 \\ & + s(t^3 - 2t^2 + t)\mathbf{P}_3 + (-2t^3 + 3t^2)\mathbf{P}_3 + s(t^3 - t^2)\mathbf{P}_4. \end{aligned} \tag{8.46}$$

For $s = 0$, Equation (8.46) becomes $(2t^3 - 3t^2 + 1)\mathbf{P}_2 + (-2t^3 + 3t^2)\mathbf{P}_3$, which can be simplified to $(3t^2 - 2t^3)(\mathbf{P}_3 - \mathbf{P}_2) + \mathbf{P}_2$. Substituting $u = 3t^2 - 2t^3$ reduces this to $u(\mathbf{P}_3 - \mathbf{P}_2) + \mathbf{P}_2$, which is the straight line from $\mathbf{P}_2$ to $\mathbf{P}_3$.

For large s, we use Equation (8.46) to calculate the mid-curve value $\mathbf{P}(0.5)$:

$$\begin{aligned} \mathbf{P}(0.5) &= \frac{s}{8}\left[(\mathbf{P}_3 - \mathbf{P}_1) + (\mathbf{P}_2 - \mathbf{P}_4)\right] + 0.5(\mathbf{P}_2 + \mathbf{P}_3) \\ &= \frac{s}{8}\left[\mathbf{P}^t(0) - \mathbf{P}^t(1)\right] + 0.5(\mathbf{P}_2 + \mathbf{P}_3). \end{aligned}$$

This is an extension of Equation (Ans.8). The first term is the difference of the two direction vectors, multiplied by $s/8$. As s grows, this term grows without limit. The second term is the midpoint of $\mathbf{P}_2$ and $\mathbf{P}_3$. Adding the two terms (a vector and a point) produces a point that may be located far away (for large s) from the midpoint, showing that the curve moves a long distance away from the start point $\mathbf{P}_2$ before changing

direction and proceeding toward the end point $\mathbf{P}_3$. Large values of s therefore feature a loose curve (low tension).

Thus, the tension of the curve can be increased by setting s close to 0 (or, equivalently, setting T close to 1); it can be decreased by increasing s (or, equivalently, decreasing T toward 0).

◇ **Exercise 8.32:** What happens when $T > 1$?

Setting $T = 0$ results in $s = 0.5$. Equation (8.45) reduces in this case to

$$\mathbf{P}(t) = (t^3, t^2, t, 1) \begin{pmatrix} -0.5 & 1.5 & -1.5 & 0.5 \\ 1 & -2.5 & 2 & -0.5 \\ -0.5 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_3 \\ \mathbf{P}_4 \end{pmatrix}, \tag{8.47}$$

a curve known as the Catmull-Rom spline. Its basis matrix is termed the parabolic blending matrix.

Example: Given the four points $(1, 0)$, $(3, 1)$, $(6, 2)$, and $(2, 3)$, we apply Equation (8.45) to calculate the cardinal spline segment from $(3, 1)$ to $(6, 2)$:

$$\begin{aligned} \mathbf{P}(t) &= (t^3, t^2, t, 1) \begin{bmatrix} -s & 2-s & s-2 & s \\ 2s & s-3 & 3-2s & -s \\ -s & 0 & s & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} (1,0) \\ (3,1) \\ (6,2) \\ (2,3) \end{bmatrix} \\ &= t^3(4s-6, 4s-2) + t^2(-9s+9, -6s+3) + t(5s, 2s) + (3, 1). \end{aligned}$$

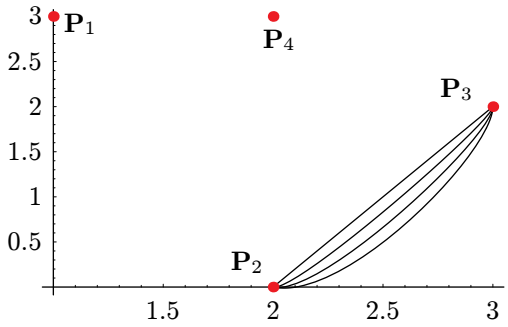
For high tension (i.e., $T = 1$ or $s = 0$), this reduces to the straight line

$$\mathbf{P}(t) = (-6, -2)t^3 + (9, 3)t^2 + (3, 1) = (3, 1)(-2t^3 + 3t^2) + (3, 1) = (3, 1)u + (3, 1).$$

For $T = 0$ (or $s = 1/2$), this cardinal spline reduces to the Catmull-Rom curve

$$\mathbf{P}(t) = (-4, 0)t^3 + (4.5, 0)t^2 + (2.5, 1)t + (3, 1). \tag{8.48}$$

Figure 8.54 shows an example of a similar cardinal spline (the points are different) with four values 0, 1/6, 2/6, and 3/6 of the tension parameter.




```
(* Cardinal spline example *)
T={t^3,t^2,t,1};
H[s_]:={{-s,2-s,s-2,s},{2s,s-3,3-2s,-s},{-s,0,s,0},{0,1,0,0}};
B={{1,3},{2,0},{3,2},{2,3}};
s=3/6; (* T=0 *)
g1=ParametricPlot[T.H[s].B,{t,0,1}];
s=2/6; (* T=1/3 *)
g2=ParametricPlot[T.H[s].B,{t,0,1}];
s=1/6; (* T=2/3 *)
g3=ParametricPlot[T.H[s].B,{t,0,1}];
s=0; (* T=1 *)
g4=ParametricPlot[T.H[s].B,{t,0,1}];
g5=Graphics[{AbsolutePointSize[4], Table[Point[B[[i]]],{i,1,4}] }];
Show[g1,g2,g3,g4,g5, PlotRange->All]
```

Figure 8.54: A Cardinal Spline Example.

8.19 Parabolic Blending: Catmull-Rom Curves

The Catmull-Rom curve (or the Catmull-Rom spline) is the special case of a cardinal spline with tension $T = 0$. This short section describes an approach to the Catmull-Rom spline where each spline segment is derived as the blend of two parabolas. This approach to the Catmull-Rom curve proceeds in the following steps:

1. Organize the points in overlapping groups of three consecutive points each. The groups are

$$[\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3], \quad [\mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4], \quad [\mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5], \quad \cdots \quad [\mathbf{P}_{n-2}, \mathbf{P}_{n-1}, \mathbf{P}_n].$$

2. Fit two parabolas, one through the first three points, $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$, and the other through the overlapping group, $\mathbf{P}_2$, $\mathbf{P}_3$, and $\mathbf{P}_4$.

3. Calculate the first curve segment from $\mathbf{P}_2$ to $\mathbf{P}_3$ as a linear blend of the two parabolas, using the two barycentric weights $1 - t$ and t.

4. Fit a third parabola, through points $\mathbf{P}_3$, $\mathbf{P}_4$, and $\mathbf{P}_5$ and compute the second curve segment, from $\mathbf{P}_3$ to $\mathbf{P}_4$, as a linear blend of the second and third parabolas.

5. Repeat until the last segment, from $\mathbf{P}_{n-2}$ to $\mathbf{P}_{n-1}$, is calculated as a linear blend of the $(n-3)$rd and the $(n-2)$nd parabolas.

Each parabola is defined by three points (which, of course, are on the same plane)

and is therefore flat. However, the two parabolas that make up the segment are not generally on the same plane, so their blend is not necessarily flat and can twist in space.

The two original parabolas are denoted by $\mathbf{Q}(u) = (u^2, u, 1)\mathbf{H}_{123}$ and $\mathbf{R}(w) = (w^2, w, 1)\mathbf{H}_{234}$, where $\mathbf{H}_{123}$ and $\mathbf{H}_{234}$ are column vectors, each depending on the three points involved. They will have to be calculated. The expression for the blended segment is $\mathbf{P}(t) = (1-t)\mathbf{Q}(u)+t\mathbf{R}(w)$. Since this expression depends on t only, we have to express parameters u and w in terms of t. We try the linear expressions $u = at + b$, $w = ct + d$.

To calculate a, b, c, and d, we write the end conditions for the two parabolas and for the curve segment (Figure 8.55a):

$$\begin{aligned}
&\mathbf{Q}(0) = \mathbf{P}_1, \quad \mathbf{Q}(0.5) = \mathbf{P}_2, \quad \mathbf{Q}(1) = \mathbf{P}_3,\\
&\mathbf{R}(0) = \mathbf{P}_2, \quad \mathbf{R}(0.5) = \mathbf{P}_3, \quad \mathbf{R}(1) = \mathbf{P}_4,\\
&\mathbf{P}(0) = \mathbf{P}_2, \qquad\qquad\quad \mathbf{P}(1) = \mathbf{P}_3.
\end{aligned}$$

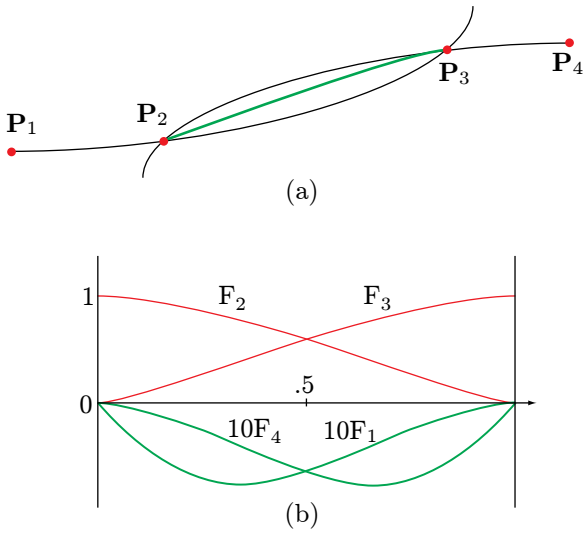


Figure 8.55: Parabolic Blending: (a) Two Parabolas. (b) The Blend Functions.

For point $\mathbf{P}_2$, we get (1) $u = 0.5$ and $t = 0$, implying $b = 0.5$, and (2) $w = 0$ and $t = 0$, implying $d = 0$. For point $\mathbf{P}_3$, we similarly get (1) $u = 1$ and $t = 1$, implying $a + b = 1 \Rightarrow a = 0.5$, and (2) $w = 0.5$ and $t = 1$, implying $c = 0.5$. This results in $u = (1 + t)/2$ and $w = t/2$.

Therefore, for the first parabola, we get

$$\left.\begin{aligned}
\mathbf{Q}(0) &= \mathbf{P}_1 = (0, 0, 1)\mathbf{H}_{123},\\
\mathbf{Q}(0.5) &= \mathbf{P}_2 = (1/4, 1/2, 1)\mathbf{H}_{123},\\
\mathbf{Q}(1) &= \mathbf{P}_3 = (1, 1, 1)\mathbf{H}_{123}.
\end{aligned}\right\} \Longrightarrow \begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix} = \begin{pmatrix}0 & 0 & 1\\ 1/4 & 1/2 & 1\\ 1 & 1 & 1\end{pmatrix}\mathbf{H}_{123} = \mathbf{M}\mathbf{H}_{123},$$

This can be solved for $\mathbf{H}_{123}$

$$\mathbf{H}_{123} = \mathbf{M}^{-1}\begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix} = \begin{pmatrix}2 & -4 & 2\\ -3 & 4 & -1\\ 1 & 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}.$$

So the first parabola is

$$\mathbf{Q}(u) = (u^2, u, 1)\mathbf{M}^{-1}\begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}.$$

The second parabola is obtained similarly:

$$\mathbf{R}(w) = (w^2, w, 1)\mathbf{M}^{-1}\begin{pmatrix}\mathbf{P}_2\\ \mathbf{P}_3\\ \mathbf{P}_4\end{pmatrix}.$$

The first curve segment is therefore

$$\begin{aligned}
\mathbf{P}(t) &= (1-t)\mathbf{Q}(u) + t\mathbf{R}(w)\\
&= (1-t)(u^2, u, 1)\mathbf{M}^{-1}\begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix} + t(w^2, w, 1)\mathbf{M}^{-1}\begin{pmatrix}\mathbf{P}_2\\ \mathbf{P}_3\\ \mathbf{P}_4\end{pmatrix}\\
&= (1-t)(2u^2 - 3u + 1, -4u^2 + 4u, 2u^2 - u)\begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\end{pmatrix}\\
&\quad + t(2w^2 - 3w + 1, -4w^2 + 4w, 2w^2 - w)\begin{pmatrix}\mathbf{P}_2\\ \mathbf{P}_3\\ \mathbf{P}_4\end{pmatrix}\\
&= (-0.5t^3 + t^2 - 0.5t)\mathbf{P}_1 + (1.5t^3 - 2.5t^2 + 1)\mathbf{P}_2\\
&\quad + (-1.5t^3 + 2t^2 + 0.5t)\mathbf{P}_3 + (0.5t^3 - 0.5t^2)\mathbf{P}_4 \qquad (8.49)\\
&= (t^3, t^2, t, 1)\begin{pmatrix}-0.5 & 1.5 & -1.5 & 0.5\\ 1 & -2.5 & 2 & -0.5\\ -0.5 & 0 & 0.5 & 0\\ 0 & 1 & 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\\ \mathbf{P}_3\\ \mathbf{P}_4\end{pmatrix}\\
&= (t^3, t^2, t, 1)\mathbf{B}\mathbf{P}, \qquad (8.50)
\end{aligned}$$

where $\mathbf{B}$ is the parabolic blending matrix.

The other segments are calculated similarly. Note that in practice, there is no need to calculate the parabolas. The program simply executes a loop where in each iteration, it uses Equation (8.49) with the next group of points to calculate the next segment.

The Catmull-Rom curve starts at point $\mathbf{P}_2$ and ends at $\mathbf{P}_{n-1}$. To make it pass through all n points $\mathbf{P}_1, \ldots, \mathbf{P}_n$, we add two more points $\mathbf{P}_0$ and $\mathbf{P}_{n+1}$. In practice, we normally select them as $\mathbf{P}_0 = \mathbf{P}_1$ and $\mathbf{P}_{n+1} = \mathbf{P}_n$. The first group of points is now $\mathbf{P}_0, \ldots, \mathbf{P}_3$, and the last one is $\mathbf{P}_{n-2}, \ldots, \mathbf{P}_{n+1}$. This also makes the method more

interactive, since two more points can be repositioned to edit the shape of the curve. The curve can also be closed, if the first and last points are set to identical values.

Equation (8.49) gives the representation of the Catmull-Rom curve in terms of the four blending functions

$$F_1(t) = (-0.5t^3 + t^2 - 0.5t), \quad F_2(t) = (1.5t^3 - 2.5t^2 + 1),$$

$$F_3(t) = (-1.5t^3 + 2t^2 + 0.5t), \quad F_4(t) = (0.5t^3 - 0.5t^2).$$

Note how F_1 and F_4 are negative (Figure 8.55b), how F_2 and F_3 are symmetric, and how the four functions are barycentric.

◇ **Exercise 8.33:** Prove the first-order continuity of the parabolic curve.

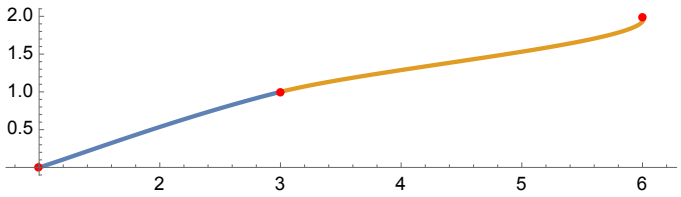
Example: Given the five points $(1,0)$, $(3,1)$, $(6,2)$, $(2,3)$, and $(1,4)$, we calculate the Catmull-Rom curve from $(1,0)$ to $(1,4)$. The first step is to add two more points, one on each end. We simply duplicate each of the two endpoints, ending up with seven points. The first segment is (from Equation (8.49))

$$\begin{aligned}\mathbf{P}_1(t) &= (-0.5t^3 + t^2 - 0.5t)(1,0) + (1.5t^3 - 2.5t^2 + 1)(1,0) \\ &\quad + (-1.5t^3 + 2t^2 + 0.5t)(3,1) + (0.5t^3 - 0.5t^2)(6,2) \\ &= (-0.5t^3 + 1.5t^2 + t + 1, -0.5t^3 + t^2 + 0.5t).\end{aligned}$$

This segment goes from point $(1,0)$ (for $t = 0$) to point $(3,1)$ (for $t = 1$). The next segment, from $(3,1)$ to $(6,2)$, is similarly

$$\begin{aligned}\mathbf{P}_2(t) &= (-0.5t^3 + t^2 - 0.5t)(1,0) + (1.5t^3 - 2.5t^2 + 1)(3,1) \\ &\quad + (-1.5t^3 + 2t^2 + 0.5t)(6,2) + (0.5t^3 - 0.5t^2)(2,3) \\ &= (-4,0)t^3 + (4.5,0)t^2 + (2.5,1)t + (3,1).\end{aligned}$$

This is identical to Equation (8.48). The curves are displayed in Figure 8.56. Calculating the other two segments is left as an exercise.




```
p1[t_]:={-0.5t^3+1.5t^2+t+1 , -0.5t^3+t^2+0.5t};
p2[t_]:={-4,0}t^3+{4.5,0}t^2+{2.5,1}t+{3,1};
g1=ParametricPlot[{p1[t],p2[t]},{t,0,1}];
g2=ListPlot[{{1,0},{3,1},{6,2}},PlotStyle->{PointSize[Medium],Red}];
Show[g1,g2,AspectRatio->.25]
```

Figure 8.56: A Catmull-Rom Spline Example.

8.20 Bézier Approximation

The previous exploration of polynomial and spline interpolations assumes that a set of data points is given and a curve should be designed that passes through the points and is also pleasing to the eye, satisfying both the curve designer and any boss/client. This section is devoted to the Bézier approximation, a different approach to curve design, where the input is a set of control points, not data points. A Bézier curve segment starts at the first point, ends at the last point, but does not pass through the interior points. Instead, these points pull the segment toward them one by one, and act as sources of attraction. The segment is attracted to each of those points in turn, as it proceeds from the first to the last point. A complete curve consists of several Bézier connected segments that form a continuous curve.

Those familiar with spline methods will easily recognize the main advantage of the Bézier approach. This approach is highly interactive, because the curve designer is free to move the control points until the perfect curve is achieved. This is why the title of this section is Bézier approximation and not Bézier interpolation.

Because of their popularity, much research has been done on Bézier methods, and there currently is a vast amount of literature and software that helps understand the principles. This section concentrates on the basics and covers only the original approach taken by Bézier himself, which is based on the Bernstein polynomials.

Pierre Etienne Bézier (pronounced "Bez-yea" or "bez-ee-ay") was an applied mathematician with the French car manufacturer Renault. In the early 1960s, encouraged by his employer, he began searching for ways to automate the process of designing cars. His methods have been the basis of the modern field of computer aided geometric design (CAGD), a field with practical applications in many areas.

Perhaps the best first reference, for those interested in Bézier methods, is [Bezier methods 24], which includes an extensive bibliography. Reference [pomax 20) is an intriguing, interactive, free web book by Mike Kamermans. For personal information about Bézier, see [Bezier 24].

8.21 The Bézier Curve

The Bézier curve is a parametric curve $\mathbf{P}(t)$ that is a polynomial function of the parameter t. The degree of the polynomial depends on the number of points used to define the curve. The method employs control points and produces an approximating curve. The curve does not pass through the interior points but is attracted by them (however, see Exercise 8.39 for an exception). It is as if the points exert a pull on the curve. Each point influences the direction of the curve by pulling it toward itself, and that influence is strongest when the curve gets nearest the point. Figure 8.57 shows some examples of cubic Bézier curve segments, where the green lines indicate the control polygon of each segment. Such a segment is defined by four points and is a cubic polynomial. Notice that one segment has a cusp and another one has a loop. The fact that a Bézier segment does not pass through the points implies that the points are not "set in stone" and can be moved. This makes it easy to edit, modify and reshape the curve, which is one reason

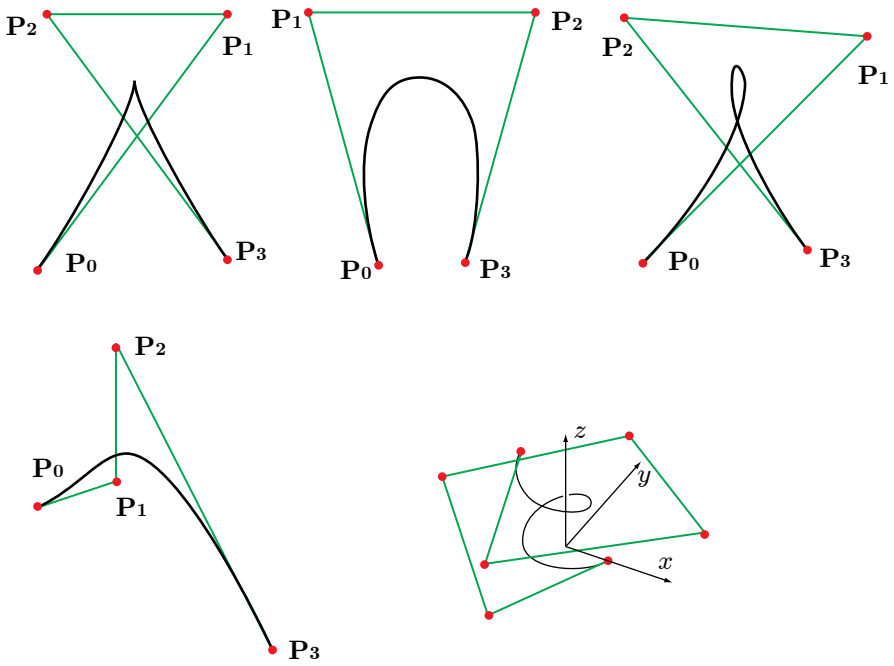


Figure 8.57: Cubic Bézier Curves and Their Control Points and Polygons.

for its popularity. The curve can also be edited by adding new points or deleting existing points.

The control polygon of the Bézier curve is the polygon obtained when the control points are connected, in their natural order, with straight segments.

How does one go about deriving such a curve? The approach shown here is based on a weighted sum, but another, equivalent approach to the same curve, based on linear interpolation, is discussed in many references.

8.21.1 Pascal Triangle and the Binomial Theorem

The Pascal triangle and the binomial theorem are related because both employ the same numbers. The Pascal triangle is an infinite triangular matrix of numbers that is constructed from the edges inside

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
\cdots & & & & \cdots & & & & \cdots & &
\end{array}
$$

We first fill the left and right edges with 1's, then compute each interior element as the sum of the two elements directly above it. As can be expected, it is not hard

to obtain an explicit expression for the general element of the Pascal triangle. We first number the rows from 0 starting at the top, and the columns from 0 starting on the left. A general element is denoted by $\binom{i}{j}$. We then observe that the top two rows (corresponding to $i = 0, 1$) consist of 1's and that every other row can be obtained as the sum of its predecessor and a shifted version of its predecessor. For example,

$$\begin{array}{rccccc} & 1 & 3 & 3 & 1 & \\ + & & 1 & 3 & 3 & 1 \\ \hline & 1 & 4 & 6 & 4 & 1 \end{array}$$

This shows that the elements of the triangle satisfy

$$\begin{aligned} &\binom{i}{0} = \binom{i}{i} = 1, \qquad i = 0, 1, \ldots, \\ &\binom{i}{j} = \binom{i-1}{j-1} + \binom{i-1}{j}, \qquad i = 2, 3, \ldots, \quad j = 1, 2, \ldots, (i-1). \end{aligned}$$

From this it is easy to derive the explicit expression

$$\begin{aligned} \binom{i}{j} &= \binom{i-1}{j-1} + \binom{i-1}{j} \\ &= \frac{(i-1)!}{(j-1)!(i-j)!} + \frac{(i-1)!}{j!(i-1-j)!} \\ &= \frac{j(i-1)!}{j!(i-j)!} + \frac{(i-j)(i-1)!}{j!(i-j)!} \\ &= \frac{i!}{j!(i-j)!}. \end{aligned}$$

Thus, the general element of the Pascal triangle is the well-known *binomial coefficient*

$$\binom{i}{j} = \frac{i!}{j!(i-j)!}.$$

The binomial coefficient is one of Newton's many contributions to mathematics. His binomial theorem states that

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}. \tag{8.51}$$

This equation can be written in a symmetric way by denoting $j = n - i$. The result is

$$(a+b)^n = \sum_{i,j \geq 0}^{i+j=n} \frac{(i+j)!}{i!j!} a^i b^j, \tag{8.52}$$

from which we can easily guess the *trinomial theorem* (which is used in the triangular Bézier surfaces)

$$(a+b+c)^n = \sum_{i,j,k\geq 0}^{i+j+k=n} \frac{(i+j+k)!}{i!j!k!} a^i b^j c^k. \tag{8.53}$$

> Bezier invented some curves
> That he used to approximate swerves
> If you use them just right
> They'll fit very tight
> And save wear and tear on your nerves
> —Darel Hardy.

8.22 The Bernstein Form of the Bézier Curve

The original approach to the Bézier curve expresses it as a weighted sum of the points. Each control point is multiplied by a weight and the products are added. We denote the control points by $\mathbf{P}_0$ through $\mathbf{P}_n$ (the number of points is therefore $n+1$) and the weights by B_i. The weighted sum that makes up a Bézier segment is

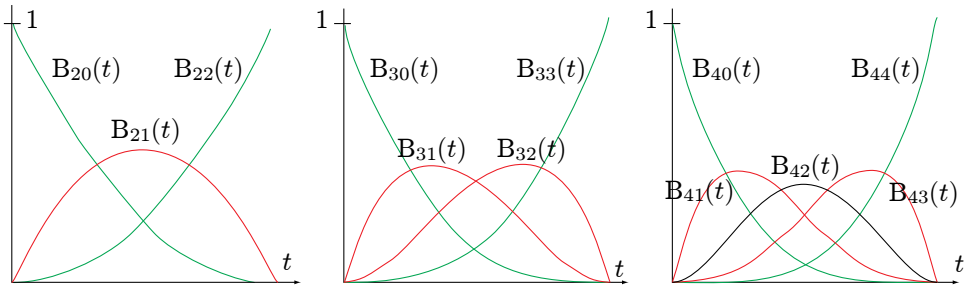
$$\mathbf{P}(t) = \sum_{i=0}^{n} \mathbf{P}_i B_i, \quad 0 \leq t \leq 1.$$

The result, $\mathbf{P}(t)$, depends on the parameter t. Since the points are given by the user, they are independent of t, so it is the weights that must depend on t. We therefore denote them by $B_i(t)$. How should $B_i(t)$ behave as a function of t?

We first examine $B_0(t)$, the weight associated with the first point $\mathbf{P}_0$. We want that point to affect the curve mostly at the beginning, i.e., when t is close to 0. Thus, as t grows toward 1 (i.e., as the curve moves away from $\mathbf{P}_0$), $B_0(t)$ should drop down to 0. When $B_0(t) = 0$, the first point no longer influences the shape of the curve.

Next, we turn to $B_1(t)$. This weight function should start small, should have a maximum when the curve approaches the second point $\mathbf{P}_1$, and should then start dropping until it reaches zero. A natural question is: When (for what value of t) does the curve reach its closest approach to the second point? A little thinking shows that it depends on the number of points. For three points ($n = 2$), the Bézier curve passes closest to the second point (the interior point) when $t = 0.5$. For four points, the curve is nearest the second point when $t = 1/3$. It is now clear that the weight functions must also depend on n and we therefore denote them by $B_{n,i}(t)$. Hence, $B_{3,1}(t)$ should start at 0, have a maximum at $t = 1/3$, and go down to 0 from there. Figure 8.58 shows the desired behavior of $B_{n,i}(t)$ for $n = 2$, 3, and 4. The five different weights $B_{4,i}(t)$ have their maxima at $t = 0$, 1/4, 1/2, 3/4, and 1.

The functions chosen by Bézier were originally derived by Sergeĭ Natanovich Bernshteĭn in 1912, as part of his work on approximation theory. They are known as the




```
(* Just the base functions bern. Note how "pwr" handles 0^0 *)
Clear[pwr,bern];
pwr[x_,y_]:=If[x==0 && y==0, 1, x^y];
bern[n_,i_,t_]:=Binomial[n,i]pwr[t,i]pwr[1-t,n-i] (* t^i x (1-t)^(n-i) *)
Plot[Evaluate[Table[bern[5,i,t], {i,0,5}]], {t,0,1}];
```

Figure 8.58: The Bernstein Polynomials for $n = 2, 3, 4$.

Bernstein polynomials and are defined by

$$B_{n,i}(t) = \binom{n}{i} t^i (1-t)^{n-i}, \quad \text{where} \quad \binom{n}{i} = \frac{n!}{i!(n-i)!} \tag{8.54}$$

are the binomial coefficients. These polynomials feature the desired behavior and have a few more useful properties that are discussed here. (In calculating the curve, we assume that the quantity 0^0, which is normally undefined, equals 1.)

The Bézier segment is now defined as

$$\mathbf{P}(t) = \sum_{i=0}^{n} \mathbf{P}_i B_{n,i}(t), \text{ where } B_{n,i}(t) = \binom{n}{i} t^i (1-t)^{n-i} \text{ and } 0 \le t \le 1. \tag{8.55}$$

Each control point (a pair or a triplet of coordinates) is multiplied by its weight, which is in the range $[0, 1]$. The weights act as *blending functions* that blend the contributions of the different points.

Here is *Mathematica* code to compute and plot the Bernstein polynomials and the Bézier curve:

```
(*Just the base functions bern.Note how "pwr" handles 0^0*)
Clear[pwr,bern,n,i,t]
pwr[x_,y_]:=If[x==0&&y==0,1,x^y];
bern[n_,i_,t_]:=Binomial[n,i]pwr[t,i]pwr[1-t,n-i]
(*t^i*(1-t)^(n-i)*)
Plot[Evaluate[Table[bern[5,i,t],{i,0,5}]],{t,0,1}]
Clear[i,t,pnts,pwr,bern,bzCurve,g1,g2];
(*Cubic Bezier curve
 either read points from file
 pnts=ReadList["DataPoints",{Number,Number}];*)
 or enter them explicitly*)
pnts={{0,0},{.7,1},{.3,1},{1,0}};
```

```
(*4 points for a cubic curve*)
pwr[x_,y_]:=If[x==0&&y==0,1,x^y];
bern[n_,i_,t_]:=Binomial[n,i]pwr[t,i]pwr[1-t,n-i]
bzCurve[t_]:=Sum[pnts[[i+1]]bern[3,i,t],{i,0,3}]
g1=Graphics[{Red, AbsolutePointSize[6],
 Table[Point[pnts[[i]]],{i,1,4}]}];
g2=ParametricPlot[bzCurve[t],{t,0,1}];
Show[g1,g2,PlotRange->All]
```

Next is similar code for a three-dimensional Bézier curve. It was used to draw the space curve of Figure 8.57.

```
Clear[pnts,pwr,bern,bzCurve,g1,g2,g3];
(*General 3D Bezier curve*)
pnts={{1,0,0},{0,-3,0.5},{-3,0,0.75},{0,3,1},
 {3,0,1.5},{0,-3,1.75},{-1,0,2}};
n=Length[pnts]-1;
pwr[x_,y_]:=If[x==0&&y==0,1,x^y];
bern[n_,i_,t_]:=Binomial[n,i]pwr[t,i]pwr[1-t,n-i]
 (*t^i x (1-t)^(n-i)*)
bzCurve[t_]:=Sum[pnts[[i+1]]bern[n,i,t],{i,0,n}];
g1=ParametricPlot3D[bzCurve[t],{t,0,1},DisplayFunction->Identity];
g2=Graphics3D[{AbsolutePointSize[2],Map[Point,pnts]}];
g3=Graphics3D[{AbsoluteThickness[2],
(*control polygon*)
 Table[Line[{pnts[[j]],pnts[[j+1]]}],{j,1,n}]}];
g4=Graphics3D[{AbsoluteThickness[1.5],
(*the coordinate axes*)
 Line[{{0,0,3},{0,0,0},{3,0,0},{0,0,0},{0,3,0}}]}];
Show[g1,g2,g3,g4,AspectRatio->Automatic,PlotRange->All,Boxed->False]
```

◇ **Exercise 8.34:** Design a heart-shaped Bézier curve based on nine control points.

When Bézier started searching for such functions in the early 1960s, he set the following requirements [Bézier 86]:

1. The functions should be such that the curve passes through the first and last control points.

2. The tangent to the curve at the start point should be $\mathbf{P}_1 - \mathbf{P}_0$, i.e., the curve should start at point $\mathbf{P}_0$ moving toward $\mathbf{P}_1$. A similar property should hold at the last point.

3. The same requirement is generalized for higher derivatives of the curve at the two extreme endpoints. Hence, $\mathbf{P}^{tt}(0)$ should depend only on the first point $\mathbf{P}_0$ and its two neighbors $\mathbf{P}_1$ and $\mathbf{P}_2$. In general, $\mathbf{P}^{(k)}(0)$ should only depend on $\mathbf{P}_0$ and its k neighbors $\mathbf{P}_1$ through $\mathbf{P}_k$. This feature provides complete control over the continuity at the joints between separate Bézier curve segments.

4. The weight functions should be symmetric with respect to t and $(1-t)$. This means that a reversal of the sequence of control points would not affect the shape of the curve.

5. The weights should be barycentric, to guarantee that the shape of the curve is independent of the coordinate system.

6. The entire curve lies within the convex hull of the set of control points.

The definition listed in Equation (8.55), using Bernstein polynomials as the weights, satisfies all these requirements. In particular, requirement 5 is proved when Equation (8.51) is written in the form $[t+(1-t)]^n = \cdots$. Following are the explicit expressions of these polynomials for $n = 2$, 3, and 4.

Example: For $n = 2$ (three control points), the weights are

$$\begin{aligned} B_{2,0}(t) &= \tbinom{2}{0} t^0 (1-t)^{2-0} = (1-t)^2, \\ B_{2,1}(t) &= \tbinom{2}{1} t^1 (1-t)^{2-1} = 2t(1-t), \\ B_{2,2}(t) &= \tbinom{2}{2} t^2 (1-t)^{2-2} = t^2, \end{aligned}$$

and the curve is

$$\begin{aligned} \mathbf{P}(t) &= (1-t)^2\mathbf{P}_0 + 2t(1-t)\mathbf{P}_1 + t^2\mathbf{P}_2 \\ &= \left((1-t)^2, 2t(1-t), t^2\right)(\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2)^T \\ &= (t^2, t, 1)\begin{pmatrix} 1 & -2 & 1 \\ -2 & 2 & 0 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{P}_0 \\ \mathbf{P}_1 \\ \mathbf{P}_2 \end{pmatrix}. \end{aligned} \tag{8.56}$$

This is the quadratic Bézier curve.

◇ **Exercise 8.35:** Given three points $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$, calculate the parabola that goes from $\mathbf{P}_1$ to $\mathbf{P}_3$ and whose start and end tangent vectors point in directions $\mathbf{P}_2 - \mathbf{P}_1$ and $\mathbf{P}_3 - \mathbf{P}_2$, respectively.

In the special case $n = 3$, the four weight functions are

$$\begin{aligned} B_{3,0}(t) &= \tbinom{3}{0} t^0 (1-t)^{3-0} = (1-t)^3, \\ B_{3,1}(t) &= \tbinom{3}{1} t^1 (1-t)^{3-1} = 3t(1-t)^2, \\ B_{3,2}(t) &= \tbinom{3}{2} t^2 (1-t)^{3-2} = 3t^2(1-t), \\ B_{3,3}(t) &= \tbinom{3}{3} t^3 (1-t)^{3-3} = t^3, \end{aligned}$$

and the curve is

$$\mathbf{P}(t) = (1-t)^3\mathbf{P}_0 + 3t(1-t)^2\mathbf{P}_1 + 3t^2(1-t)\mathbf{P}_2 + t^3\mathbf{P}_3 \tag{8.57}$$

$$\begin{aligned} &= \left[(1-t)^3, 3t(1-t)^2, 3t^2(1-t), t^3\right]\left[\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3\right]^T \\ &= \left[(1-3t+3t^2-t^3), (3t-6t^2+3t^3), (3t^2-3t^3), t^3\right]\left[\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3\right]^{\mathbf{T}} \\ &= (t^3, t^2, t, 1)\begin{pmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 3 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{P}_0 \\ \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_3 \end{pmatrix}. \end{aligned} \tag{8.58}$$

It is clear that $\mathbf{P}(t)$ is a cubic polynomial in t. It is the cubic Bézier curve. In general, the Bézier curve for points $\mathbf{P}_0$, $\mathbf{P}_1$,..., $\mathbf{P}_n$ is a polynomial of degree n.

◇ **Exercise 8.36:** Given the curve $\mathbf{P}(t) = (1 + t + t^2, t^3)$, find its control points.

◇ **Exercise 8.37:** The cubic curve of Equation (8.58) is drawn when the parameter t varies in the interval $[0, 1]$. Show how to substitute t with a new parameter u such that the curve will be drawn when $-1 \le u \le +1$.

◇ **Exercise 8.38:** Calculate the Bernstein polynomials for $n = 4$.

◇ **Exercise 8.39:** Generally, the Bézier curve passes through the first and last control points, but not through the intermediate points. Consider the case of three points $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ on a straight line. Intuitively, it seems that the curve will be a straight line and would therefore pass through the interior point $\mathbf{P}_1$. Is that so?

> Equations are just the boring part of mathematics. I attempt to see things in terms of geometry.
>
> —Stephen Hawking.

8.23 Linear Perspective

Today, the term "perspective" is commonly used. It has several meanings and is used in several contexts. A typical dictionary definition of this term has the following parts:

- Perspective is a technique for drawing a three-dimensional scene on a two-dimensional surface such that when viewed from a particular point, the viewer gets the right impression of their height, width, depth, and relative positions.
- Perspective is also a point of view or an attitude toward something.

The type of perspective covered here is referred to as linear. Linear perspective is a well-defined concept and is easy to treat mathematically. In contrast, the term nonlinear perspective is ill-defined and may refer to different approaches, techniques, and variations of linear perspective. Some examples are: (1) An image drawn on the surface of a sphere or a cylinder. (2) An image stretched or sheared in complex ways. (3) A very wide scene with overlapping regions, each with its own vanishing points. (4) Instead of straight lines converging to a point, have curves that converge to the same vanishing point (5) Fisheye perspective, where the central part of the image is blown more than the rest. (6) Mercator Perspective, where the rules of Mercator Mapping are applied to an image.

In his 2001 book, *Secret Knowledge: Rediscovering the Lost Techniques of the Old Masters*, as well as in the accompanying 2002 BBC documentary *David Hockney: Secret Knowledge*, Hockney mentions perspective only in passing. He is interested in the progress made in accuracy and realism in Western art by artists using optical devices such as a curved mirror, camera obscura, camera lucida, and lenses.

Hockney describes an experiment he and his helpers did in support of his theory. They copied hundreds of paintings and drawings and arranged them on a long wall chronologically (Figure 8.59). The works were arranged from the oldest (around 1300) on

the left, to the newest (around 1750) to the right. They were also arranged geographically from Southern Europe on the bottom to Northern Europe on top. It was then easy to notice that a revolution has occurred in art around 1420, and that it started in Italy. Artworks before 1420 were all flat, while works after 1420 show how linear perspective entered into European art, and gradually propagated from south (Italy, Spain, and France) to north (Flanders, Holland, and Germany).

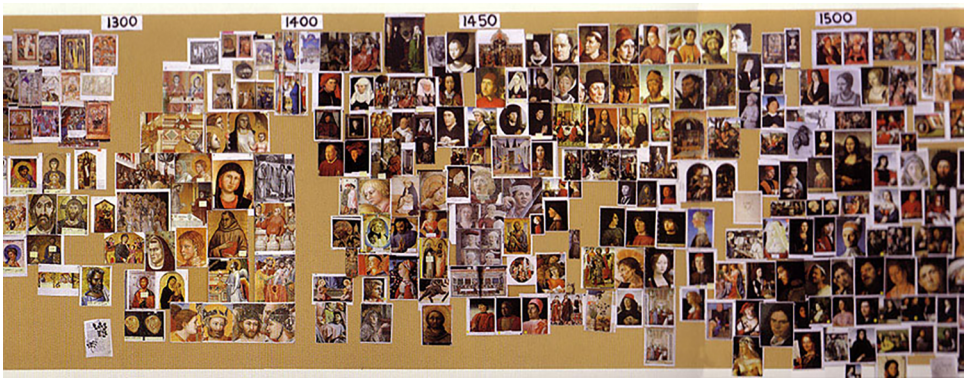


Figure 8.59: Hockney's Great Wall.

Hockney refers to paintings and drawings made after 1420 as "correct" and "almost like a photograph," but when we look at those works, we don't immediately see the effects of lenses and mirrors. Instead, we feel that they look real because they use correct perspective. Groups of lines that we know are parallel are seen to converge to vanishing points. Distant objects look small; they are foreshortened.

At time stamp 34:30 in part 1 of his documentary, Hockney mentions perspective. He defines it as "perspective is an abstraction. A device for putting what we see in three-dimensions onto a two-dimensional surface. It gives us an illusion of space."

A quick Internet search verifies that many readers of Hockney's book, as well as viewers of the BBC documentary contested his theory of the use of optical devices in art. See, for example, reference [Kreutz 01]. The critics presented many claims and arguments, but since this text is about perspective, not mirrors and lenses, I would like to limit it to just one interesting hint on how to recognize pictures drawn or painted with the use of a camera obscura. The hint is, look for an abundance of left hands.

It is generally agreed that only 10% of the population is left handed, and not much is known about left-handedness in monkeys. (See also Exercise 7.14.) Figure 8.60 shows the little-known painting *The licentious kitchen maid* (c. 1665) by Pieter van Roestraten, which depicts such a scene and thereby suggests that the painter started by tracing a camera obscura image of the scene. The other painting is the 1628 *Laughing Boy with a Wine-Glass*, by Frans Hals.

8.23.1 Pioneers of Perspective

Regardless of Hockney's claim about optical aids, his great wall full of historical artworks should be enough to convince anyone that the use of linear perspective started around





Figure 8.60: Paintings Containing Dominant Left Hands.

1420, which raises the question of precedence. Who was the first to come up with the concept of linear perspective? It is generally agreed that this person was Filippo Brunelleschi, and that he already had a clear picture of perspective in his mind as early as 1413.



Brunelleschi was an Italian architect, designer, goldsmith and sculptor; one of many contemporary renaissance men. Among his many other achievements, he is considered a founding father of Renaissance architecture, and is recognized as the first modern engineer, planner, and sole construction supervisor. Adding to all this is the fact that in 1421 Brunelleschi became the first person in the Western world to receive a patent. His greatest achievement is designing the dome of the Florence Cathedral and supervising its construction, a mammoth project that took 17 years.

We continue with the contributions made to the understanding of perspective by three Renaissance figures, Brunelleschi, Masaccio, and Alberti.

Brunelleschi

Filippo Brunelleschi, known to his contemporaries as "Pippo," was born in Florence in 1377. His father, Ser Brunellesco di Lippo Lapi, was a prosperous notary, but young Filippo showed an interest in machines and in solving mechanical problems. (The term

"ser" was a title of respect, while "di Lippo Lapi" indicates that Brunelleschi's grandfather was named Lippo and was from the Lapi family.) Filippo was therefore apprenticed, at age 15, to a local goldsmith. For the next six years he learned to cast metals, work with enamel, engrave and emboss silver, and use precious metals to decorate manuscripts with gold leaf and to make jewels and religious artifacts.

After completing his apprenticeship in 1398 at age 21, Brunelleschi was sworn as a master goldsmith and became a well-known goldsmith in Florence and other cities. From 1401 to 1416 or 1417, he seems to have spent most of his time in Rome (although this is uncertain), working as a goldsmith, making clocks, and surveying the many ruins of the eternal city. Returning to Florence after 13 years of absence, Brunelleschi, then 40, became involved in the competition for the great dome of the Santa Maria del Fiore Cathedral. This was to be both the largest dome ever attempted, with a diameter of more than 143 feet, and the tallest one, starting at a height of about 170 feet off the ground and reaching about 280 feet. (The lantern on top of it adds more than 70 feet to that.)

Even though known as a goldsmith, not an architect, Brunelleschi won the 1418 competition because of his original approach to the problem. The novel aspect of his plan for the dome was to build it without any scaffolding. (The term "centering" was then used.) This idea, and the 1:12 model of the dome that he built in brick to demonstrate his method, helped convince the committee of judges to give him the commission. He then spent the years from 1420 to 1436 supervising the construction while also designing and building ingenious machines to haul heavy loads to the top.

Brunelleschi, a true Renaissance man both because of his interests and achievements and because of his time period, died in 1446. Like Donatello, Masaccio, da Vinci, and Michelangelo, he never married. For more information on Brunelleschi, his work, and his times, see [King 00], [Saalman 70], and [Walker 02].

A biography of Brunelleschi [Manetti 88] was written in the 1480s, four decades after the death of its subject, by his pupil Antonio Manetti, which brings us to Brunelleschi's contribution to perspective. In this biography, Manetti describes Brunelleschi's panel drawing, a trompe l'oeil that was then used by Brunelleschi in an experiment that fuses nature and art, similar to an optical trick. This historically-important painting has since been lost, but it, and the experiment, are described in detail by Manetti.

Trompe l'oeil (French for "deceiving the eye," pronounced "tromp loy").
1. A style of painting that gives an illusion of photographic reality.
2. A painting or effect created in this style.

The peepshow experiment. Brunelleschi placed himself at a point three braccia (about six feet) inside the doorway of the not yet completed cathedral of Santa Maria del Fiore. His idea was to specify a precise viewing point at which a viewer could compare a real scene with a perspective painting of the same scene. Looking outside across the Piazza del Duomo, he clearly saw, about 115 feet away, the Baptistery of San Giovanni, one of Florence's most familiar landmarks. This structure was a good choice for the study of perspective because it is shaped like an octagon, so someone standing in front of it sees its three front walls in two-point perspective. (It also features left–right symmetry, so reflecting it horizontally does not change its shape.)

Brunelleschi then painted what he saw through the doorframe—the Baptistery and

some of the surrounding streets—in perspective on a small panel about 12 inches wide, and with a burnished silver strip at its top. Finally, he drilled a small hole in the panel at the center of the Baptistery's eastern doors (the ones facing the cathedral, Figure 8.61a) because this point of the Baptistery would be directly opposite the eye of a viewer standing at the specified viewing point.

> The world having so long been without artists of lofty soul or inspired talent, heaven ordained that it should receive from the hand of Filippo the greatest, the tallest, and the finest edifice of ancient and modern times, demonstrating that Tuscan genius, although moribund, was not yet dead.
>
> —Giorgio Vasari, *The Lives of the Artists* (1567).

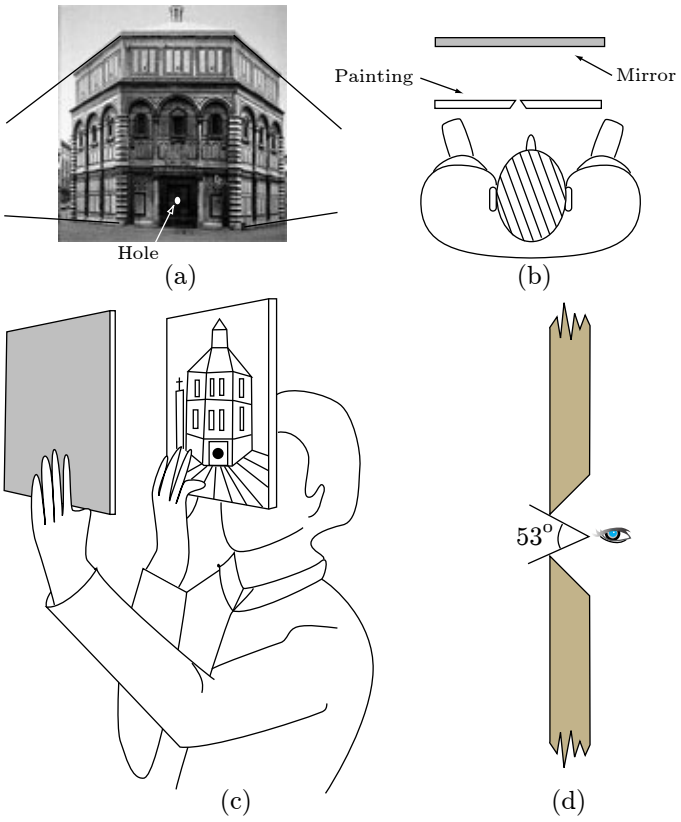


Figure 8.61: Brunelleschi's Experiment in Perspective.

Brunelleschi then rotated the panel 180° and looked through the hole at the Baptistery. He then inserted a mirror (a rare and expensive item in Italy at that time) and held it at arm's length as shown in Figure 8.61bc and looked at his painting reflected in the mirror (where the silver top reflected the sky and clouds). This became

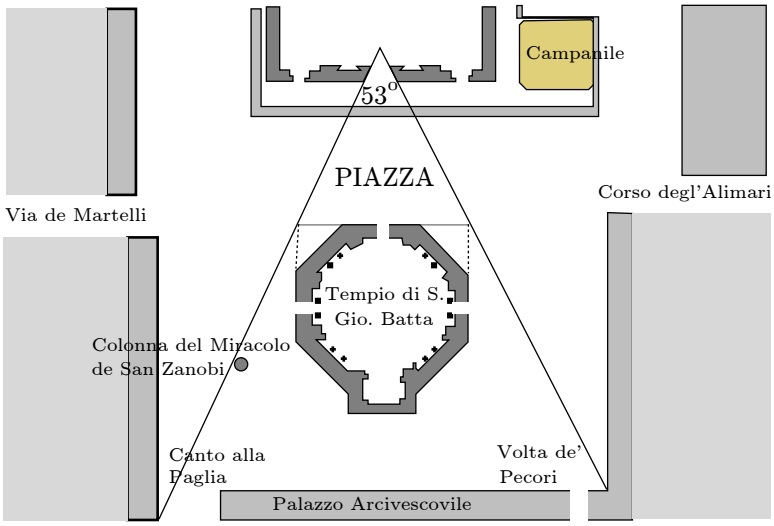


Figure 8.62: Plan of the Piazza del Duomo, Florence (After [Sgrilli 33]).

Brunelleschi's celebrated peepshow experiment, which demonstrated the lifelike qualities of linear perspective.

In his biography, Manetti claims to have held this painting in his hands and to have repeated the experiment. He was unable to tell the difference between the image reflected in the mirror and the real scene, observed without the mirror. (However, modern travelers to Florence recommend the use of a pair of heavy-duty tripods to hold the image and the mirror at their precise locations.)

[Brunelleschi] had made a hole in the panel on which there was this painting; ... which hole was as small as a lentil on the painting side of the panel, and on the back it opened pyramidally, like a woman's straw hat, to the size of a ducat or a little more. And he wished the eye to be placed at the back, where it was large, by whoever had it to see, with the one hand bringing it close to the eye, and with the other holding a mirror opposite, so that there the painting came to be reflected back ... which on being seen, ... it seemed as if the real thing was seen: I have had the painting in my hand and have seen it many times in these days, so I can give testimony.

—Antonio Manetti, *The Life of Brunelleschi* (1480s).

Manetti must have seen this panel many times, and it was also listed among the personal effects of Lorenzo the Magnificent after his death in 1492. After that point, all traces of it have disappeared.

Manetti mentions another interesting fact. The painting was about 12 inches wide and Brunelleschi recommended watching it from a distance of six inches, so the reflection seen in the mirror appears to be at a distance of 12 inches from the viewer. We know that $\tan 26.6^\circ = 0.5$, which implies that the apex angle of an isosceles triangle whose

height equals its base is $2 \times 26.6 \approx 53°$. This trigonometric fact suggests that, as seen from the viewing point specified by Brunelleschi, the Baptistery spans a viewing angle of about 53°, and this is verified by Figure 8.62, which follows the site plan given by [Sgrilli 33]. Finally, Manetti mentions that the diameter of the hole on the painted side of the panel was about the thickness of a bean (6–7 mm). Figure 8.61d illustrates how the same angle of 53° is obtained if the eye of the viewer is glued to the back of the panel (where according to Manetti the hole was bigger, about the size of a ducat, 20 mm) and the thickness of the panel is the same 6–7 mm.

While this experiment of Brunelleschi's is well-known and is mentioned and discussed in many places (including youtube), Manetti mentions a second experiment—this time with two-point perspective, illustrated in Figure 8.63—that was simpler and did not include a mirror. Here is a quotation, from [Edgerton 75]:

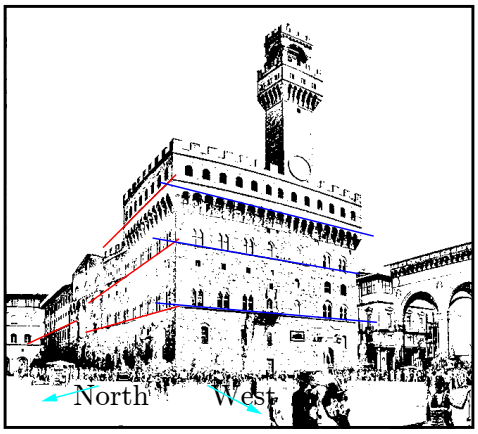


Figure 8.63: Brunelleschi's Second Experiment.

He made in perspective the piazza of the palace of the Signori of Florence, with everything on it and round about it, as much as can be seen, standing outside the piazza or really on a level with it, along the facade of the Church of Santo Romolo, past the corner of Calimala Francesca, which rises on the aforesaid piazza, a few braccia towards Orto Santo Michele, whence is seen the palace of the Signori, in such a way that two faces are seen completely, that which is turned towards the West and that which is turned toward the North: so that it is a wonderful thing to see what appears, together with all the things that the view includes in that place.

Afterwards Paolo Uccello and other painters did it, who wished to counterfeit and imitate it; of which I have seen more than one, and it was not as well done as that. Here it might be said: why did he not make this picture, being of perspective, with that hole for the eye, like the little panel from the Duomo towards Santo Giovanni? This arose, because the panel of so great a piazza needed to be so big to put in it so many different things, that it could

> not, like the Santo Giovanni, be held up to the face with one hand, nor the mirror with the other; for the arm of a man is not of sufficient length that with the mirror in his hand he could hold it at its distance opposite the point, nor so strong, that he could support it.
>
> He left it to the direction of the onlooker as happens in all other paintings of all other painters, although the onlooker may not always be discerning. And in the place where he put the burnished silver in that of Santo Giovanni, here he left a void, which he made from the buildings up: and betook himself with it to look at it in a place where the natural air showed itself from the buildings upwards.

This little-known, interesting, but not very readable text is found on page 128 of [Edgerton 75], who adds the following useful information:

> This translation has been adapted from John White, *Birth and Rebirth of Pictorial Space* (London, 1968), pp. 114, 116, 117. For a more readable if less literal translation, see Saalman, *The Life of Brunelleschi*, pp. 42–46.

Today, after hundreds of years of research into Brunelleschi and his work, Manetti is not the only source. On page 125 of [Edgerton 75] I found the following surprise:

> Antonio Averlino, called Filarete, in his own treatise on architecture written during the early 1460's:
>
> "And so I believe that Pippo di Ser Brunellescho the Florentine found the way to make this plan [linear perspective] which truly was a subtle and beautiful thing, which he discovered through considering what a mirror shows to you."

Masaccio

After his 1413 historical experiments with perspective, nothing more is heard about Brunelleschi's interest in perspective. It took more than a decade for his ideas and techniques to propagate to other parts of Italy, and another decade to appear, in the early 1430's, in Northern European art.

Art historians believe that the first great Renaissance painter to use Brunelleschi's ideas and techniques in serious works of art was Tommaso di ser Giovanni di Mone (or Tommaso di ser Giovanni cassai), known to us as Masaccio, a nickname that can be translated as Big Thomas, Rough Thomas, Clumsy Thomas, Sloppy Thomas, Bad Thomas, or even Messy Thomas. He died in 1428, at age 27, and in his last two years he painted two important pictures (1) A 1425 fresco, today titled *Trinity* (or *Holy Trinity*), in the Church of Santa Maria Novella in Florence, and (2) The *Tribute Money*, a fresco in the Brancacci chapel, also in Florence (Santa Maria del Carmine). The accurate execution of one-point perspective in the former creates the illusion of a sculpture placed in a cavity in the wall, although the picture is flat. Figure 8.64 shows how the painter placed the vanishing point exactly five feet nine inches above the floor, to have it at the viewing height of an average person. Hence, a visitor to Santa Maria Novella is in the same position as Brunelleschi was when he peered across the Piazza del Duomo at a point in the center of the Baptistery western doors.

This large picture (approximately 6.7×3.2 m, or 21 ft 10.5 in by 10 ft 5 in) has a sad history of incompetent restoration and a 19th century attempt to cut it off the wall

and move it to another wall in the same Church.

> The architectural setting of this fresco [the *Trinity*] is so accurate in its perspective and so Brunelleschian in style that some scholars have suggested Brunelleschi drew the sinopia, or cartoon, on the wall for Masaccio to paint. This is certainly possible, but it is also quite possible that Masaccio—a master draftsman as well as an inspired painter—could have done the whole work himself. Perhaps it doesn't matter. The important fact for the future of Western art is that Masaccio met Brunelleschi and gained such a deep knowledge of perspective that he set a standard for every painter to follow.
>
> —Paul Robert Walker, *The Feud that Sparked the Renaissance* (2002).

The *Tribute Money*, painted by Masaccio with the help of Masolino around 1427, illustrates more of Brunelleschi's ideas. The two painters knew about the importance of the horizon line and the rule of horizon line isocephaly. They, perhaps for the first time in the history of Western art, placed the central group of figures with their heads aligned on that line, regardless of their distance from the viewer. In addition to their use of linear perspective, the painters applied the following techniques to add depth and realism to the painting:

- Horizon line isocephaly. When a group of objects is located roughly along a horizontal line, align their heads along that line. Distant objects should appear smaller and with their lowest points raised. Figure 8.65 (*Delivery of the Keys*, by Pietro Perugino) is a clear example of this rule. Notice especially the man marked by the green arrow, who is positioned slightly behind his neighbor.

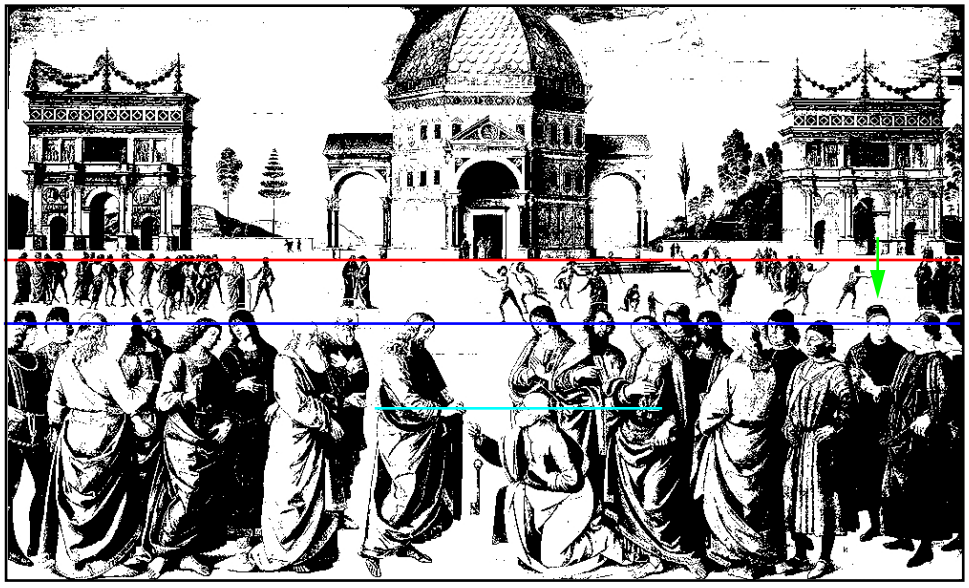


Figure 8.65: Perugino's *Delivery of the Keys*.

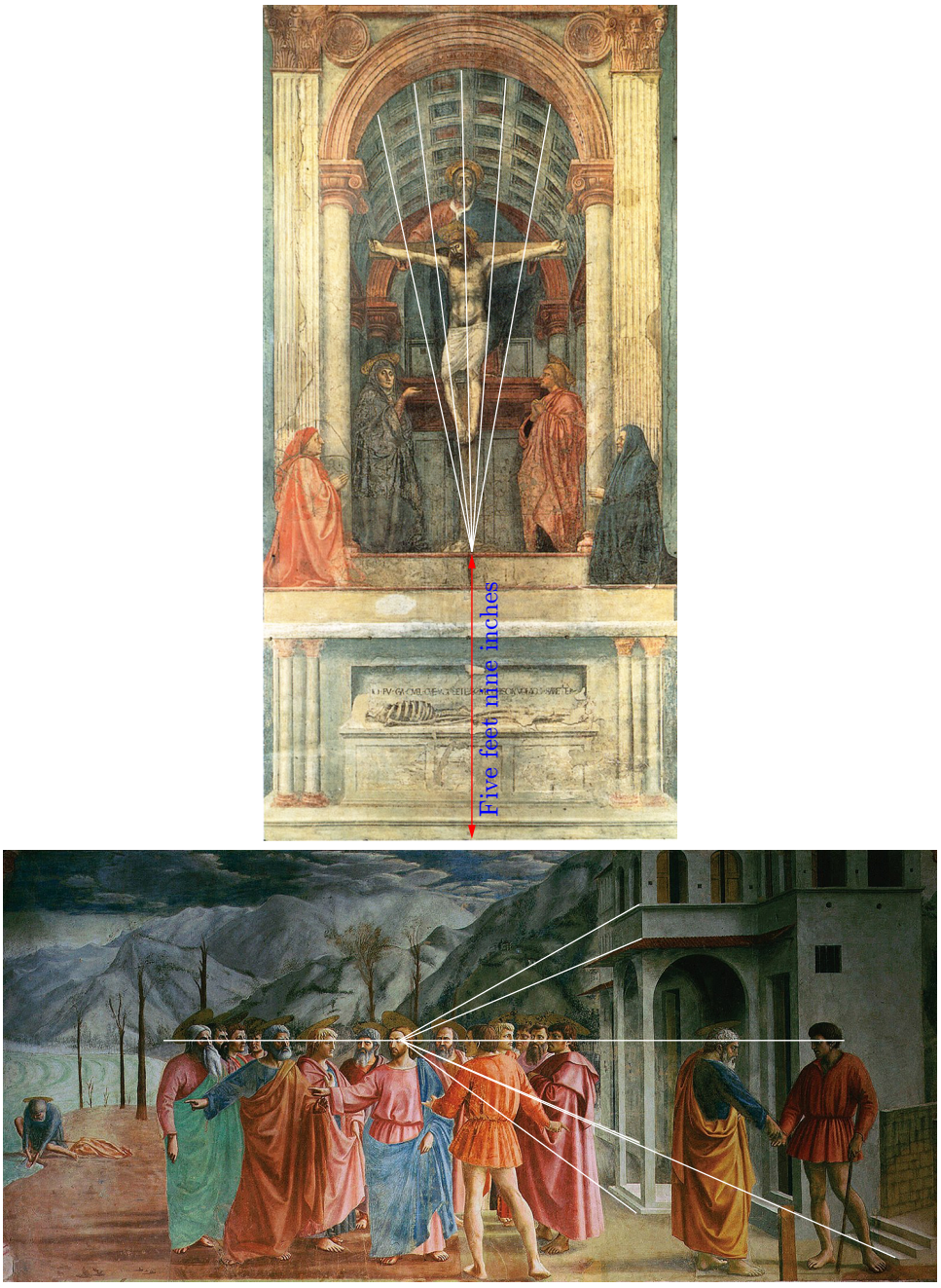


Figure 8.64: Masaccio's *Holy Trinity* and *The Tribute Money*.

- Use the minimum needed to create the illusion of depth. Masaccio has only one, minor structure on the right of the picture, providing only a few short lines to locate the vanishing point and create the desired depth effect. By minimizing the importance of this structure, the painter has made the group of people at the center of the painting the dominant feature of the scene.

- Continuous narrative. This is the idea of combining multiple spatial scenes (or several moments in time) in a single picture. The painting tells the story of Christ and Peter in three temporal scenes, placed (1) at the center, (2) on the left, and (3) on the right side of the painting.

- Naturalism. The idea is that an image appears natural if any idealization or stylization is omitted. In this picture, naturalism is expressed in the lifelike expressions of the figures and in their period clothing.

- Atmospheric (or aerial) perspective. The illusion of depth is created by making distant objects appear misty, murky, lighter, and more ambiguous and undefined than nearby objects. This creates a sense of distance between the figures in the foreground and the mountains in the background.

  Atmospheric perspective was also used by Leonardo da Vinci. It is based on the perception that contrasts of color and shade appear greater in nearby objects than in distant objects, and that warm colors (such as red, orange, and yellow) appear to bring objects closer to the viewer, while cool colors (blue, violet, and green) appear to do the opposite. Aerial perspective is also used in East Asian art, where zones of mist are sometimes used to separate nearby and distant parts of the scene.

Alberti

Debates over who should have priority in science can be very worrisome and unpleasant for everyone involved. An instance of this would be the uncovering of gravity. Historians, educators, and textbooks have focused solely on Isaac Newton for 400 years, but recent discoveries indicate that Robert Hooke had a similar idea a few months earlier. Who was the one to uncover gravity, Newton or Hooke? Even if Hooke originated the idea, Newton is credited with precisely explaining gravity in mathematical terms, demonstrating its impact on falling apples and celestial bodies, and transforming it into a reputable and practical scientific discipline.

(There are reasons to believe that important historical figures such as da Vinci, Brahmagupta, and Bhaskar Acharya also came up with the basic idea of gravity long before Hooke, but similarly lacked the mathematical tools needed to place their ideas on a firm and rigorous basis.)

A similar controversy surrounds Brunelleschi and Leon Battista Alberti, concerning the discovery (or perhaps rediscovery) of perspective. The former was the genius who first came up with the main concept and the essential rules of perspective, while the latter put those rules into mathematical language and, even more important, put them on paper. In this way Alberti influenced future artists and showed how to incorporate perspective in artworks. We can therefore claim that Alberti, though not the first to understand perspective, was the person who truly transformed Italian art, and later European art, from the flat style used for millennia to the modern form. Instead of

considering him the discoverer, we can safely say that he was the chief figure in the development of the fundamental technique of perspective.

A word about their backgrounds. Brunelleschi's father was a notary who could provide his son with only elementary education. The basic abacco education at his time stressed practical subjects and lacked advanced applications of mathematics as well as the classics (Greek and Latin). Brunelleschi was also secretive. He worried about his competitors stealing his ideas, which is perhaps why he left no writings.

Alberti, on the other hand, came from a rich, patrician family, and had an excellent education. He studied the classics at the university of Padua, followed by law, mathematics, and perhaps also optics, at the university of Bologna. As a clerk, a secretary, and a confidante to the papal curia, he spent years traveling in Italy, visiting the powerful courts in places such as Ferrara, Mantua, and Urbino. In 1387 his family was banished from Florence by the Albizzi ruling family. In 1428, when the Albizzis fell from power, the Alberti family could return, and Leon Battista was finally able, in 1434, to visit the city of his ancestors. He was quick to develop relations with the Medici, to absorb the new approaches to art, and to see the Duomo, the Baptistery, the Palazzo della Signoria, the Ponte Vecchio, and other major locations.

It was at this time, in 1435–1436, while under the effects of the bold architecture and revolutionary art and ideas that he absorbed in Florence, that Alberti wrote and published (in the vernacular Italian) his great book *Il Trattato della Pittura e I Cinque Ordini Architttonici* (*On Painting*). The Latin version of this book, titled *De Pictura*, was perhaps also written by him. In this book he describes a simple geometric method for constructing a correct one-point perspective of a horizontal grid on a vertical picture plane. This method was later simplified by Piero della Francesca in his 1478 mathematical treatise *De prospectiva pingendi* (*On perspective painting*).

Alberti's approach to perspective drawing, as described in his book, is based on the concept of a window and it consists of two main steps, the vanishing point and the distance point. The first step starts by drawing a rectangle, the window through which the final scene would be viewed. Next, the painter decides how tall the main figures (plants, animals, or humans) should be. In Figure 8.66 this is denoted by H, and the artist (1) partitions the bottom of the window into equal-size parts, each about $H/3$ units wide, (2) draws the horizon line in the window at height H, and (3) establishes a vanishing point at the center of the horizon line and draws the orthogonals from the bottom to the vanishing point.

A modern artist or designer places the horizon wherever he wants the eye level of the viewer to be. To Alberti, the horizon line separates the window into two parts. The lower part contains those objects and figures that are not higher than the eye of the viewer. The upper part is the place for the large objects (often the sky, mountains, or tall buildings) that are farther away. Distant figures are drawn with their top at the horizon line, which creates the illusion of depth. The modern term for this rule is horizon line isocephaly.

> The canvas is an open window through which I see what I want to paint.
> —Leon Battista Alberti.

The second step in Alberti's perspective construction, the distance point, is illustrated in Figure 8.67.

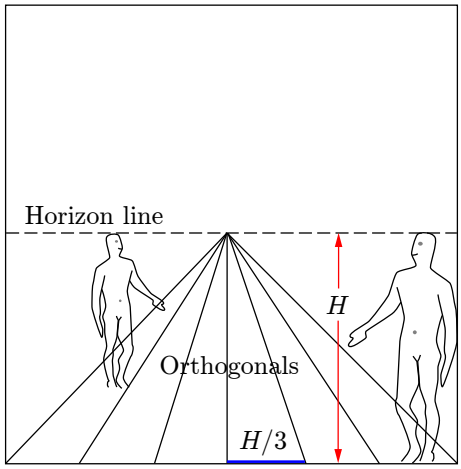


Figure 8.66: Alberti's Window.

The left part of the figure shows a side view where the picture plane is intercepted by a family of visual rays that emanate from the viewer's eye. Each ray connects the eye to one of the equally-spaced grid points on the ground. The point where the ray intercepts the picture plane is then transferred to the front view (on the right part of the figure) to indicate where to place the particular transversal in the picture. It is easy to see how the transversals, which in reality are equally spaced, get closer and closer in the perspective view. The last step is to draw a diagonal line in the front view to check for the accuracy of this geometric construction. This diagonal should intersect the corners of the squares it traverses.

The middle part of the figure illustrates the effects of the viewer getting near the picture plane. The transversals in this case are being spread out, indicating a large perspective effect (significant differences between nearby and distant objects). The bottom part of the figure shows what happens when the painter chooses a low horizon. The transversals are brought closer, the bottom part of the figure become less dominant, and there is room for more objects above the horizon. This makes sense for pictures that depict stars, flying objects (identified or not), or clouds.

Figures 8.68 and 8.69 demonstrate these concepts with photographs.

In his book, Alberti also shows how a floor, accurately drawn in perspective, can serve to determine the correct dimensions (both horizontal and vertical) of objects positioned on the floor and elsewhere in the picture. Figure 8.70 (top) illustrates how a grid on a floor is used to determine the height of a large, box-like object placed on the floor. Alberti used the braccio (plural braccia), a length unit that equals approximately 58 cm (or 23 in, roughly the length of a man's arm), and a length of four braccia, measured on the floor, is employed to determine the heights of the box at its front and back.

The bottom part of the figure shows how the system of traversals can also be applied to the ceiling, resulting in two grids of lines converging to the same vanishing point.

It is such precisely described methods and techniques that distinguish Alberti from his predecessors and justify his title as a "pioneer" or "originator" of perspective.

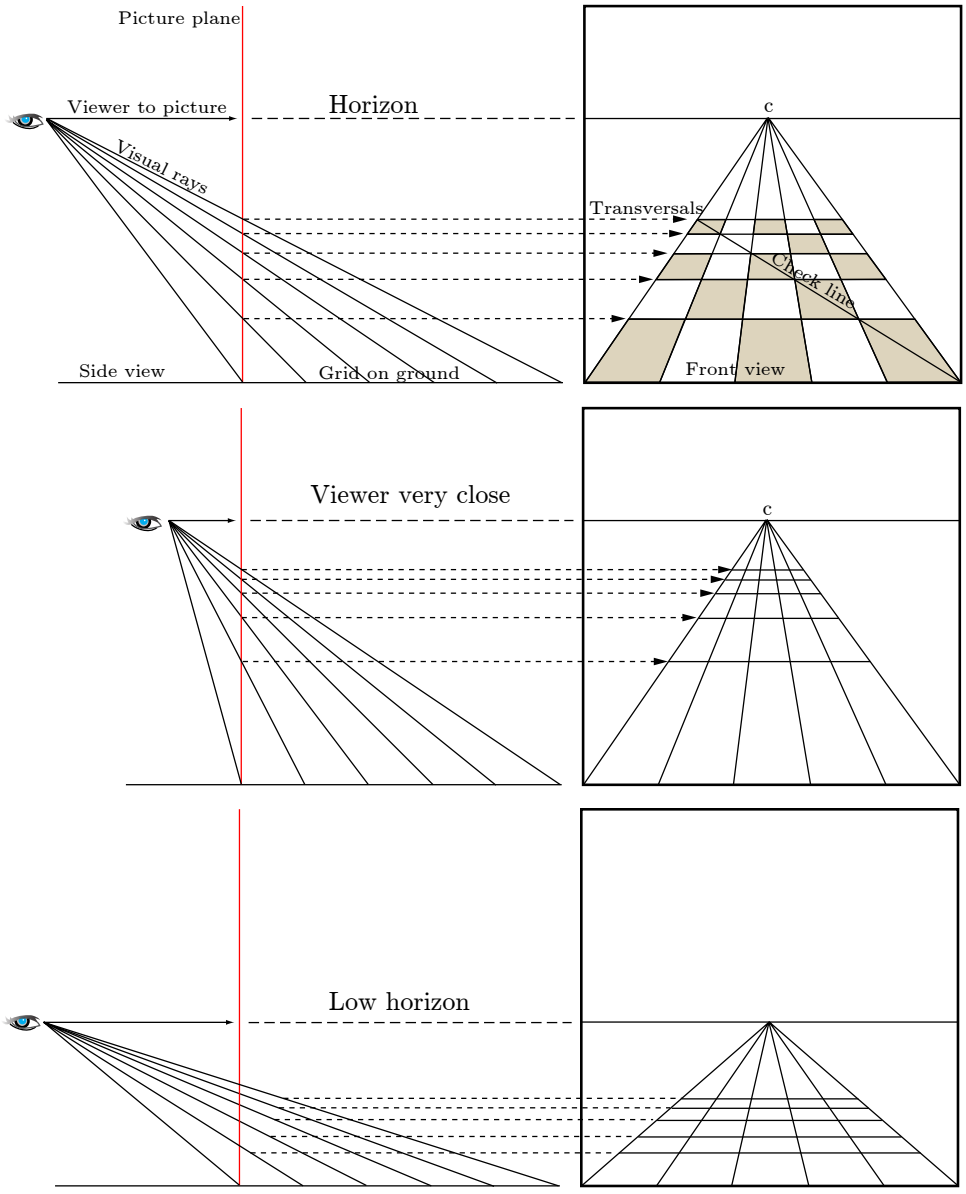


Figure 8.67: Alberti's Method of Traversals in One-Point Perspective.



Photo by Tom Barrett on unsplash



Photo by Saj Shafique on unsplash



Photo by Massimo Virgilio on unsplash



Photo by Adrian Botica on unsplash

Figure 8.68: One-Point Perspective Demonstrated.

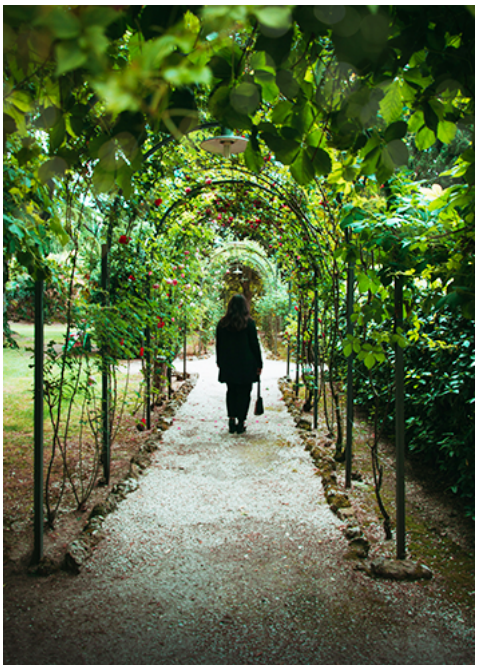


Photo by Edoardo Busti on unsplash



Photo by Daniel Norris on unsplash



Photo by Dave Lowe on unsplash



Photo by Massimo Virgilio on unsplash

Figure 8.69: One-Point Perspective Demonstrated.

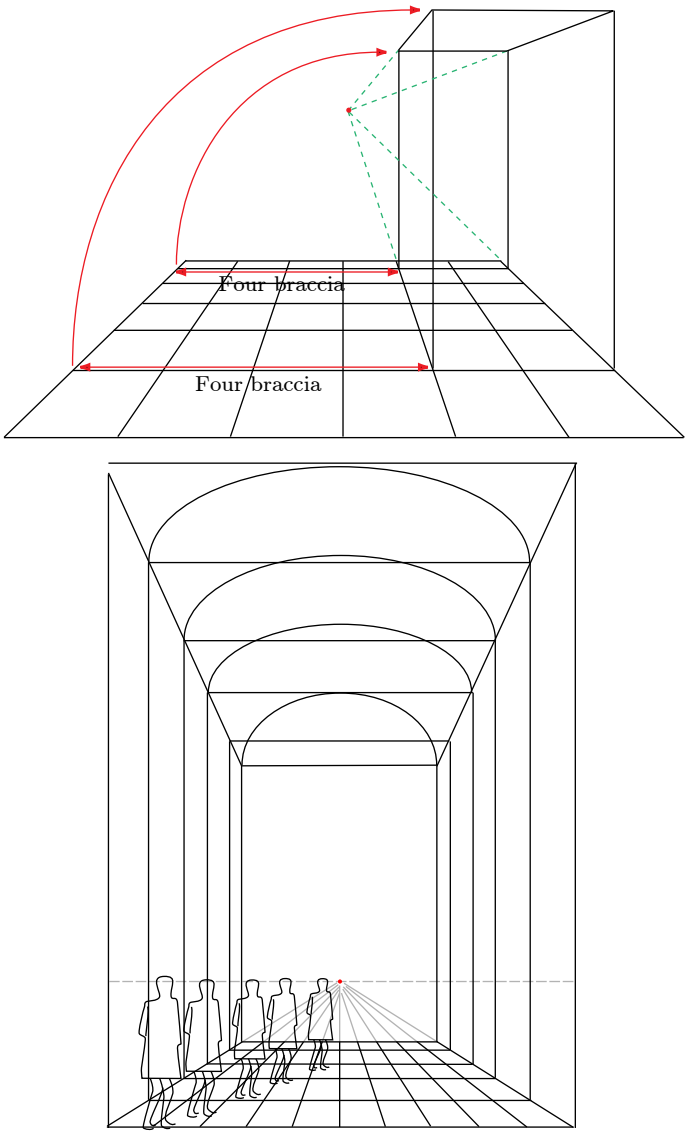


Figure 8.70: Determining Vertical Dimensions from the Floor.

◇ **Exercise 8.40:** Given the simple two-point perspective of Figure 8.71, show how the equally-spaced red vertical lines were constructed.

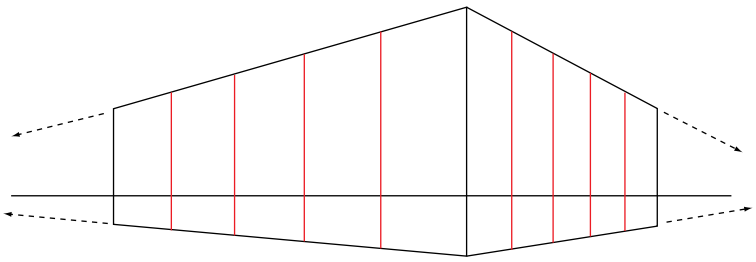


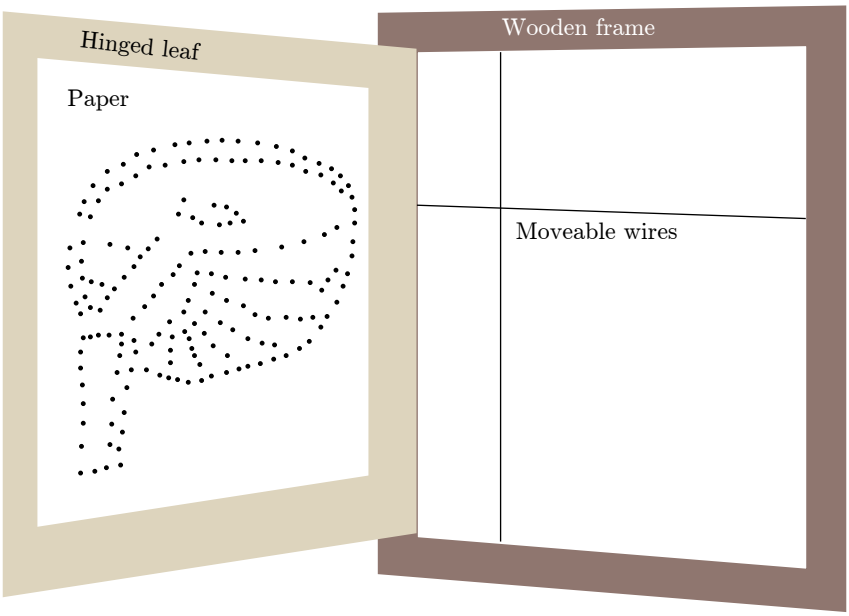
Figure 8.71: Two-Point Perspective with Equally-Spaced Lines.

> Alberti "was a man of rare brilliance, acute judgement, and extensive learning... Surely there was no field of knowledge, however remote, no discipline, however arcane, that escaped his attention."
>
> —Angelo Poliziano.

The great German painter Albrecht Dürer showed how to extend Alberti's approach to three-dimensional objects (Figure 8.72). Lay the object (a lute in the figure) on a table behind a frame and attach a string with a pulley and a weight to the wall in front of the frame. A wooden leaf is attached to the frame with hinges, and a sheet of blank paper is mounted on the leaf. Now move the free end of the string to an arbitrary point on the object and determine the point where the string intercepts the frame. (This is done by two moveable wires or threads, as shown in the upper part of the figure.) Remove the string temporarily, close the hinged leaf, and mark the intersection point of the wires on the paper. This is repeated for many points on the object, which later permits the artist to interpolate the points and complete the drawing.

In contrast with renaissance and classical artists, who mostly tried to create works true to nature, many impressionist and modern artists consider the use of color and technique more important than accurate perspective. Figures 8.73 and 8.74 show examples of this approach. Figure 8.73 shows: (1) The van Gogh chair, (2) A 13th century Madonna with Child in the Italo-Byzantine style, and (3) Altarpiece of the Sienese Cathedral by Maesta. The yellow chair was painted by Vincent van Gogh several times during his short stay in Arles. Even a quick glance may convince the viewer that something is wrong. However, van Gogh fans claim that his mastery of color, combined with his technique and style, resulted in paintings full of appeal and charm, in spite of the crude perspective (or even because of it). Many of the chairs of David Hockney are also examples of what is often referred to as divergent or inverse perspective, reverspective. Such effects can also be achieved with a hypercentric camera, whose lens makes distant objects appear larger than closer ones.

Figure 8.74 shows: (a) Fresco from the grotto of Touen Houang, China, Tang dynasty (618–906). (b) Pseudo Jacopino di Francesco, *Mary Magdalene Washing Christ's Feet*. (c) Andrei Rublev. (c. 1360–1430) *The Trinity*, late 1400's, early 1500's, State Tretyakov Gallery, Moscow. (d) Refectory of Abbazia di Pomposa (Pomposa abbey), *Miracle of the Abbot Guido Strambiati*, Giotto frescoes in Rimini from 1316–1320.



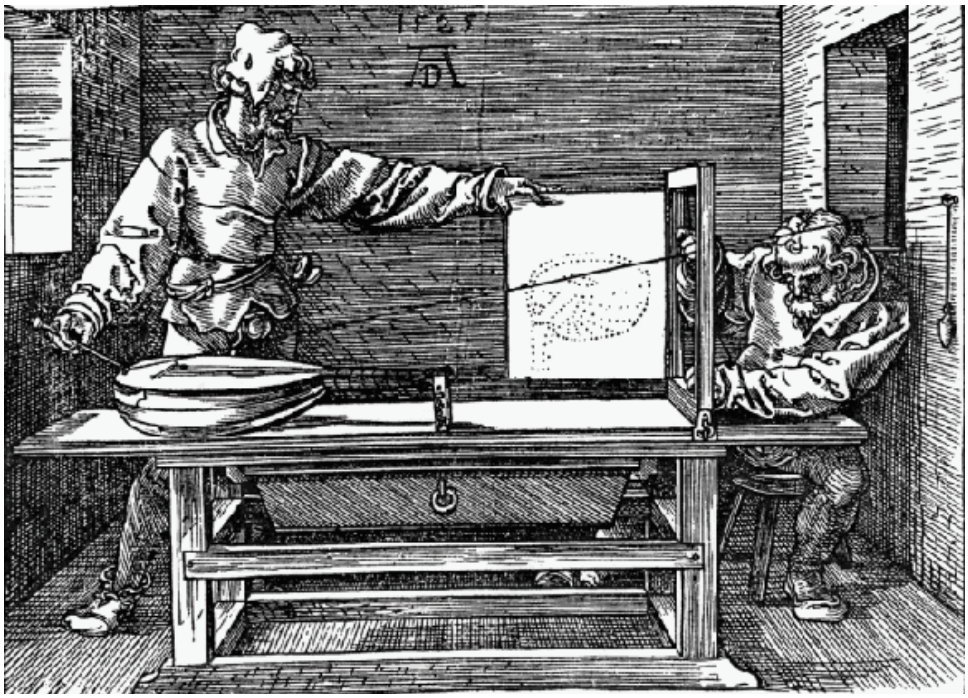


Figure 8.72: Dürer's Method of Perspective Drawing.



Figure 8.73: Examples of Divergent Perspective.

> Indoors, he [Vincent] painted a canvas of the oddly proportioned room where he slept, attempting to convey his feeling of peace there, ...
>
> —Cynthia Saltzman, *Portrait of Dr. Gachet*, 1999.



(a) (b) (c) (d)

Figure 8.74: Examples of Divergent Perspective.

8.24 Perspective: Basic Concepts

The basic ideas and rules of perspective are derived from the well-known observation that distant objects seem small. This observation, coupled with Figure 8.75, illustrates the main idea behind perspective, the concept of converging lines and vanishing points. The figure also serves to understand the main rule of vanishing points: Any group of parallel lines on an object converges to a vanishing point, except if the lines are perpendicular to the line of sight (the groups of gray lines in the figure). See Figure 8.76 for examples of several vanishing points.

Another important aspect of perspective has to do with distances. When an object is moved away from the viewer, it appears smaller, but its perspective also diminishes.

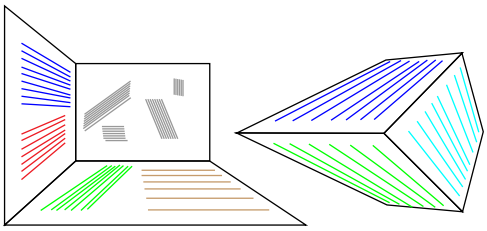
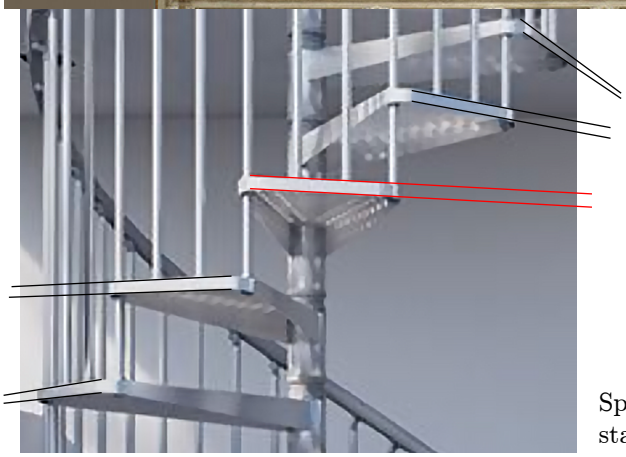


Figure 8.75: Parallel Lines Converging.



Raphael
The school of Athens



Spiral
staircase

Figure 8.76: Several Vanishing Points.

The amount of perspective seen by the viewer depends on the relation between the depth of the viewed object and its distance from the viewer. This is illustrated in Figure 8.77. Part (a) of the figure shows a cube of 10 cm on a side that is viewed from a distance of 10 cm. The cube's back face is therefore 20 cm from the viewer, twice the distance of the front face. The back face therefore seems to the viewer much smaller than the front face, and the object is seen with considerable perspective. In part (b), the cube is moved 90 cm away from the viewer. Its front face ends up at 100 cm while its back face is at 110 cm from the viewer. The difference between front and back is now much smaller compared with the distance from the viewer, causing the back face to appear only a shade smaller than the front, with the result that the object appears to have a lesser perspective (Figure 8.77b).

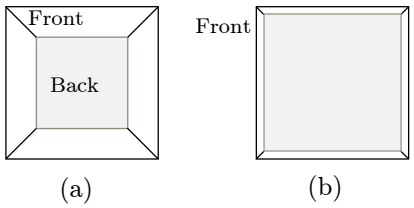


Figure 8.77: (a) More and (b) Less Perspective.

◇ **Exercise 8.41:** In addition to featuring less perspective, a distant object also looks small. Can we bring such an object closer without increasing its perspective?

The horizon line. The placement of the horizon line is one of the first decisions made when planning a picture in perspective. In principle, the artist/designer can place this line anywhere in the window, but the following discussion explains how the position of the horizon affects the overall impression that the final picture creates in the mind of the viewer.

The viewer always assumes that the horizon line is at his height (see the top part of figure 8.64), which is why placing the horizon in the middle of the picture puts the viewer in a 1:1 relationship with the subjects at its center, while also creating the feeling that higher and lower objects have secondary importance. The artist should consider the viewer a floating eye that can be placed anywhere in the picture window, and whose placement will determine the effect the rest of the image has on the viewer.

When the horizon line is placed high in the picture (Figure 8.78), the viewer feels that he is observing from a high position, perhaps standing on a hill or on a high bridge, parachuting, or flying. Such a picture shows more of the top of the objects below and less of their sides. Clouds, stars, and the bottom parts of high-flying airplanes may appear at the top of the picture. An example is a picture of a gigantic airship surrounded by small aircraft, all flying over a city with skyscrapers.

A high horizon also creates the feeling of big ground and small sky. This is the right choice when there are many objects on the ground and the painter wants them to appear less imposing, while stressing certain objects located high, near the painter himself. One example is a view from the Eiffel Tower, showing some of the steel structures near the

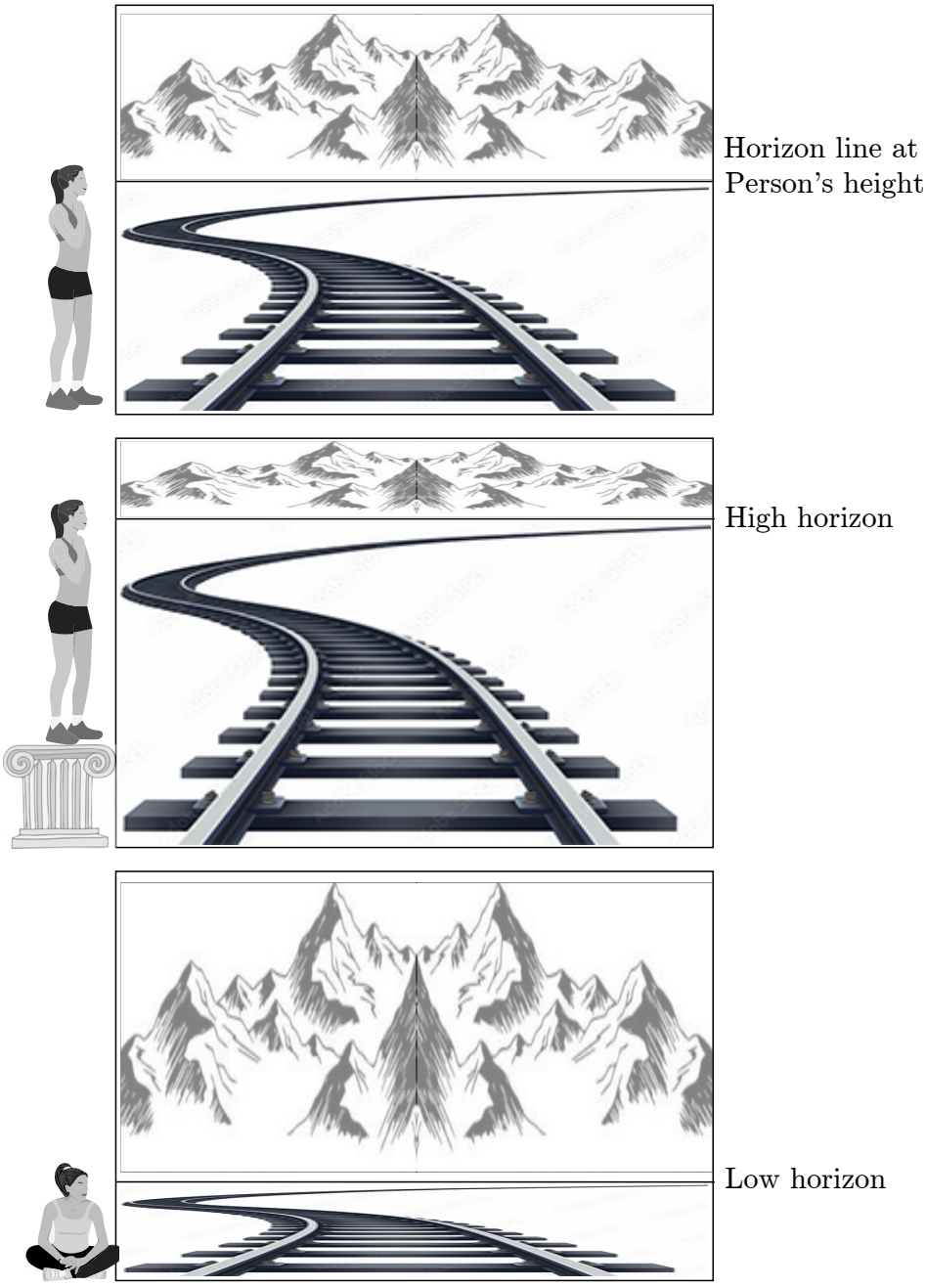


Figure 8.78: High and Low Horizons.

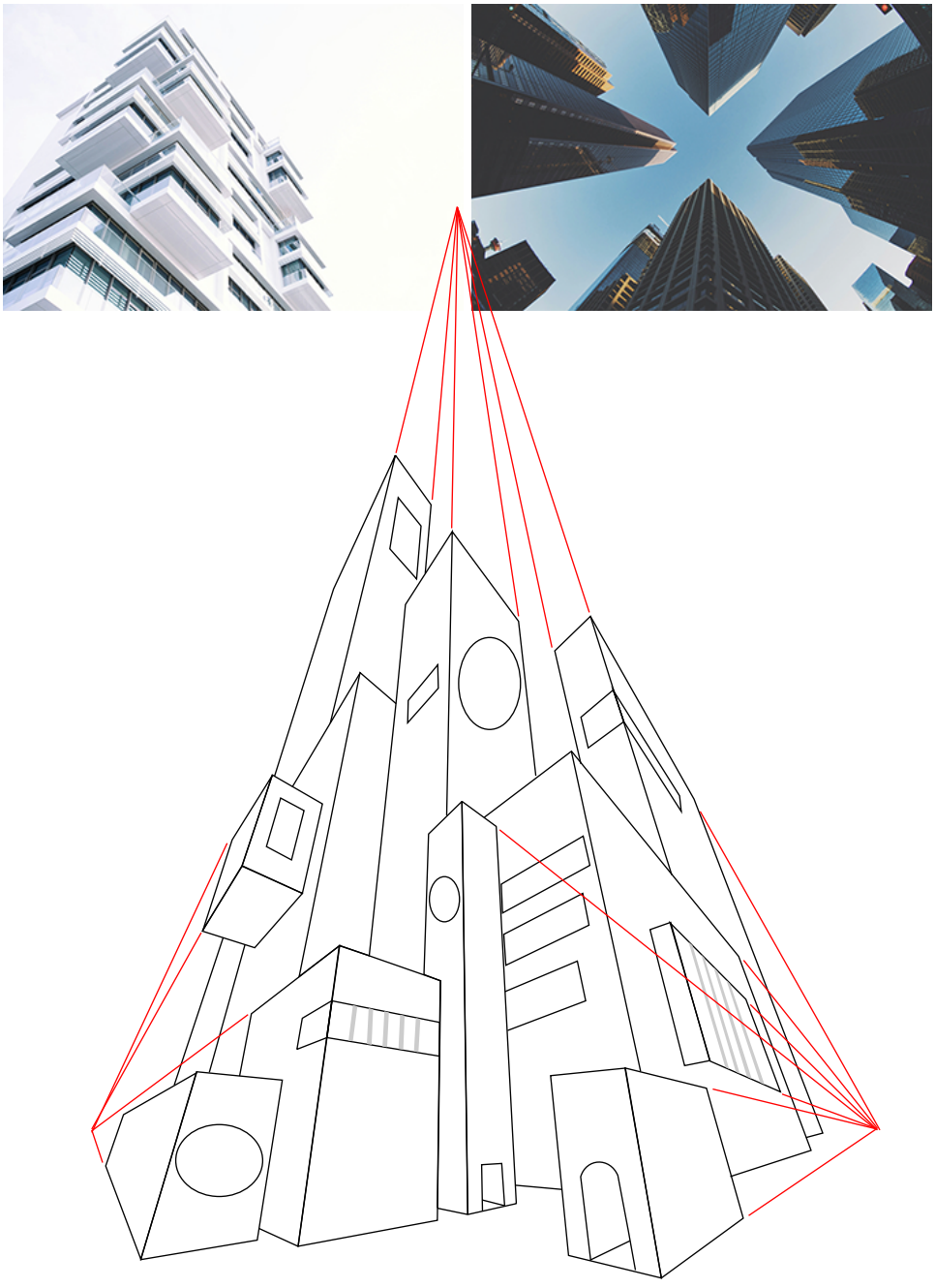


Figure 8.79: Low Horizon Examples.

painter/photographer in much detail, while the many buildings, people, and bridges on the ground are minor details. An extreme example is the 1935 woodcut *Inside St. Peter's*, by M. C. Escher.

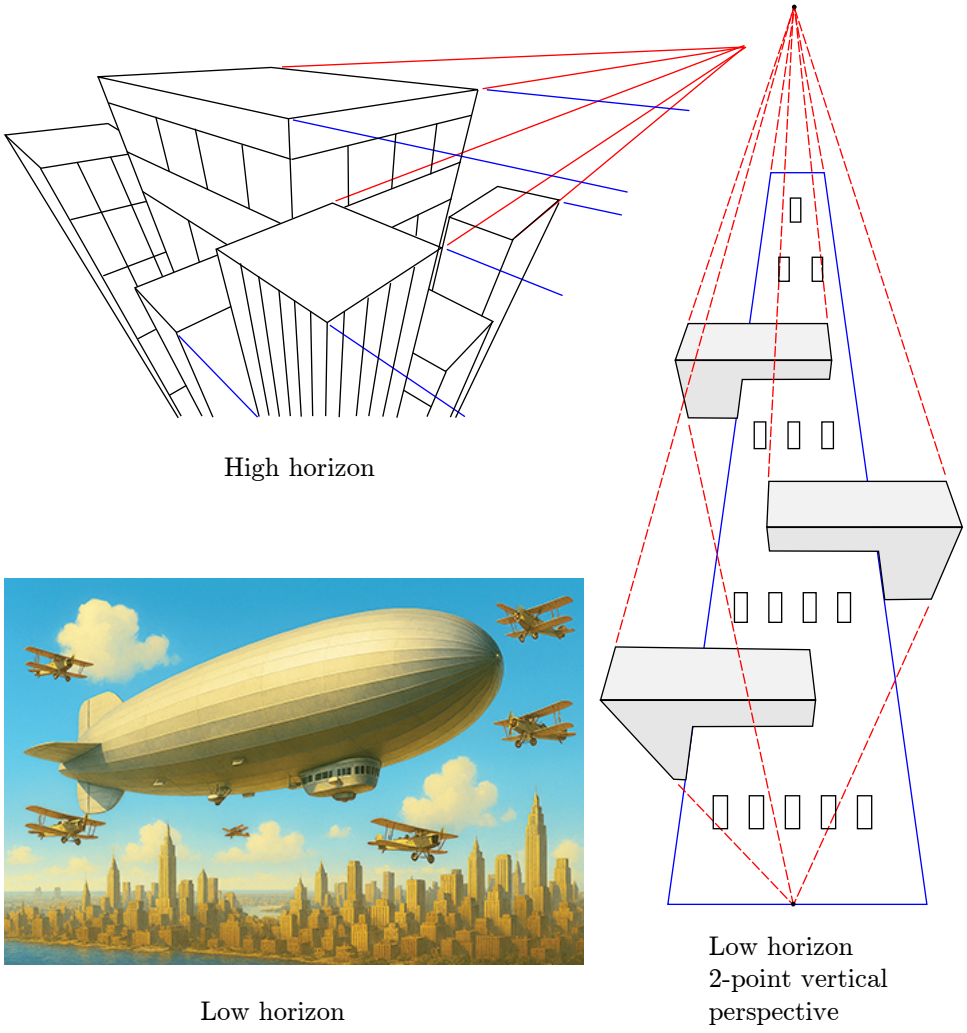


Figure 8.80: High and Low Horizons.

Figure 8.80 illustrates a few examples of low and high horizon. Especially interesting is the right-hand side of the figure, which is an example of a 2-point vertical perspective, where the two vanishing points are along a vertical line. A similar example appears in Figure 8.81, which also has images featuring high horizons.

A low horizon, such as in Figure 8.79, creates in the viewer the feeling that he is sitting, squatting, or perhaps standing low, in a hole in the ground. The sky is big and

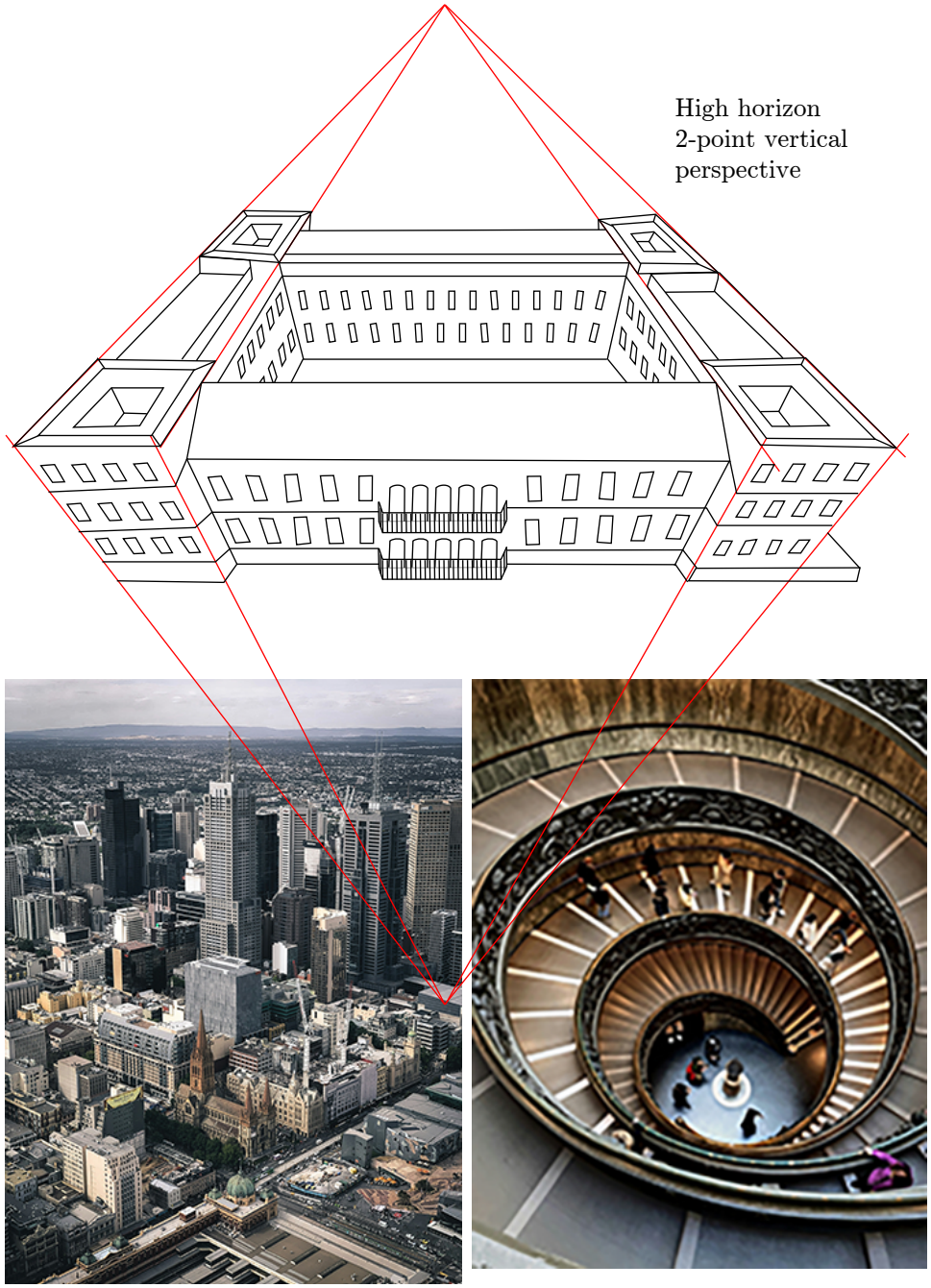
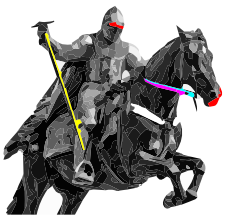


Figure 8.81: High Horizons.

the ground is small. The viewer sees details of nearby low-lying objects, but more of the front and bottom parts of higher objects. Even nearby objects on the ground may be obscured by closer objects, which creates the feeling of limited perspective. Such a placement makes sense for a picture where the bottom parts of objects are dominant.

> In a hole in the ground there lived a hobbit.
> —J. R. R. Tolkien, *The Hobbit*, 1937.



Another dramatic effect is achieved when the main objects in such a picture dominate the viewer because he is so much below them. The object may then seem heroic and larger than life. An example is a knight in armor riding a huge horse, as seen by someone sitting low. Two extreme examples of a low horizon are the 1952 woodcut *Puddle* and the 1955 lithograph *Three Worlds*, both by Escher.

Speaking of Escher, his intriguing 1947 painting *High and Low* has a low horizon in the lower half of the picture and a high horizon in its top half.

8.25 The Mathematics of Perspective

The term "projection" has meanings in a wide variety of fields including physics, chemistry, biology, linguistics, and art. In mathematics, this term is used, among others, in connection with set theory, measure theory, and relational algebra. In this text, our interest is in perspective projection, a method of transforming three-dimensional objects to a two-dimensional plane, where they satisfy the rules of perspective. For comparison, consider the much simpler parallel projection. The parallel projection of the three-dimensional point (x, y, z) on the xy plane is simply the two-dimensional point (x, y), but the perspective projection of (x, y, z) has to take into account the location and orientation of the viewer. The perspective projection seen by a distant viewer is not only small, but also features shallow perspective. This is discussed at the beginning of Section 8.24 and is illustrated in Figure 8.77.

The mathematics of perspective projection is easy to derive and to apply to various situations. The mathematical problem involves three entities, a three-dimensional object to be projected, a projection plane, and a viewer watching the projection on this plane. The object and the viewer are located on different sides of the projection plane, and the problem is to determine what the viewer will see on the plane. It is like having a transparent plane and looking through it at an object. Specifically, given an arbitrary point $\mathbf{P} = (x, y, z)$ on the object, we want to compute the coordinates (x^*, y^*) of its projection $\mathbf{P}^*$ on the projection plane. Once all the points of the object are projected and drawn on the projection plane, the object appears on the plane in perspective. Thus, the problem is to find a transformation $\mathbf{T}$ that will transform $\mathbf{P}$ to $\mathbf{P}^*$. We use the notation $\mathbf{P}^* = \mathbf{PT}$, where $\mathbf{T}$, Equation (8.62), is the as-yet-unknown 4×4 perspective transformation matrix whose elements are to be determined. This matrix is 4×4 and not 3×3 as expected, because of the use of homogeneous coordinates.

Often, there is no need to compute the projections of every object point. If $\mathbf{P}_1$ and $\mathbf{P}_2$ are the two endpoints of a straight segment on the object, only their projections $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ need be computed and a straight segment drawn between them on the plane. In the case of a curve, it is enough to compute the projections of several points on the curve and either interpolate them on the projection plane or simply connect them with short, straight segments.

It is obvious that what the viewer will see on the projection plane depends on the position and orientation of the viewer. The viewer and the object have to be located on different sides of the plane, and the viewer should look at the plane. If the viewer moves, turns, or tilts his head, he will see something else on the projection plane or may not even see this plane at all. Similarly, if the object is moved or if the projection plane is moved or is rotated, the projection would change. Thus, the mathematical expressions for perspective must depend on the location and orientation of both the viewer and the projection plane, as well as on the location of each point $\mathbf{P}$ of the object.

We start with the special case, the standard position, where the viewer is positioned at a special location, looking in a special direction at a specially-placed projection plane. There is no need to consider the orientation of the object because each point $\mathbf{P}$ on the object is projected individually. In reference [Salomon 11] this treatment is generalized and it is shown how to project an object on any projection plane and with the viewer located anywhere and looking in an arbitrary direction.

The earlier discussion of the rules of perspective and of converging lines implies that the transformation matrix $\mathbf{T}$ should satisfy the following conditions:

1. As the object is moved away from the projection plane, its projection shrinks. This is in accordance with the well-known fact that distant objects appear small.

2. The projection of a distant object features shallow perspective, as illustrated by Figure 8.77. The reader may claim that the projection of a distant object is too small to be seen, so the loss of perspective may not matter, but the point is that we can look at a distant object through a telescope. This instrument brings the object closer, so it looks big, but still features shallow perspective.

3. Any group of straight parallel lines on the object seems to converge to a vanishing point, except if the lines are perpendicular to the line of sight of the viewer.

The remainder of this section derives this special case of perspective projection in four steps as follows:

1. We describe the special case and state the rule of projection.
2. The mathematical expressions are derived using only similar triangles.
3. We show that this rule satisfies the three requirements above.
4. The general three-dimensional transformation matrix is derived from this rule. Because of the use of homogeneous coordinates this is a 4×4 matrix that can both transform the points of an object and project them on a plane.

Step 1. The special case discussed in this section places the viewer at point $(0, 0, -k)$, where k, a positive real number, is a parameter chosen by the user. The viewer looks in the positive z axis, so the line of sight is the vector $(0, 0, 1)$. Finally, the projection plane is the xy plane. In order for the projection to make sense, we state again that the viewer and the object must be located on different sides of the projection plane, and this implies that all the points of the object must have nonnegative z coor-

dinates. (The points will normally have positive z coordinates, but they may also be of the form $(x, y, 0)$; i.e., located on the projection plane itself.)

The special case discussed here is referred to as the standard position and is illustrated in Figure 8.82a. In addition to being commonly used in practice, this case simplifies the mathematical derivation. To locate the perspective projection of point $\mathbf{P}$ we compute the line segment that connects $\mathbf{P}$ to the viewer at point $(0, 0, -k)$ and place the projected point $\mathbf{P}^*$ where this segment intercepts the xy plane. (The segment always intercepts the xy plane because the object and the viewer are located on opposite sides of the plane.) Because the projection plane is the xy plane, the coordinates of the projected point are $(x^*, y^*, 0)$, indicating that it is essentially two-dimensional.

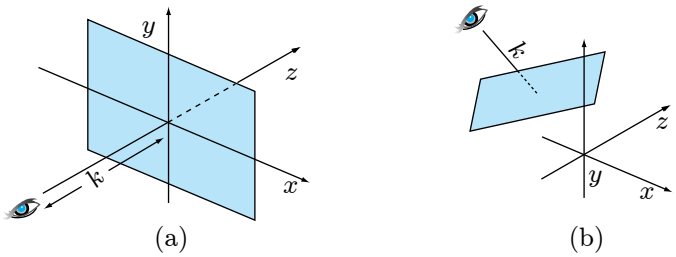


Figure 8.82: (a) Standard and (b) Nonstandard Positions.

It is important to realize that the viewer and the projection plane constitute a single unit and should be moved and rotated together. This is illustrated in Figure 8.82b and especially in Figure 8.83a, which shows the viewer-plane unit moving around the object and the viewer looking at the object from different directions, examining various projections of it on the plane. It is pointless to move the viewer around the object while the projection plane stays at the same location (Figure 8.83b) because such a viewer will generally not even be looking at the projection plane. Thus, the projection plane must move with the viewer and must remain perpendicular to the line of sight of the viewer and at a distance of k units from him (although k may be varied by the user).

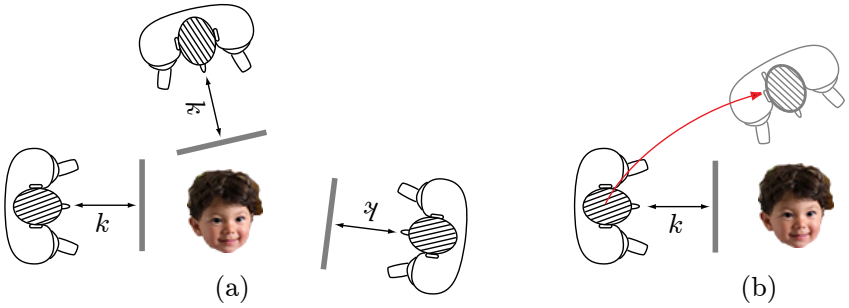


Figure 8.83: Moving the Viewer and the Projection Plane.

Step 2. The two similar triangles of Figure 8.84 yield the simple relations

$$\frac{x^*}{k} = \frac{x}{z+k} \quad \text{and} \quad \frac{y^*}{k} = \frac{y}{z+k},$$

from which we obtain

$$x^* = \frac{x}{(z/k)+1} \quad \text{and} \quad y^* = \frac{y}{(z/k)+1}. \tag{8.59}$$

(Some authors assign the x coordinate a negative sign. This is a result of the difference between left-handed and right-handed coordinate systems.) The $+1$ in the denominator of Equation (8.59) is important. It guarantees that the denominator will never be zero. The denominator can be zero only if $z/k = -1$, but k is positive and z is nonnegative.

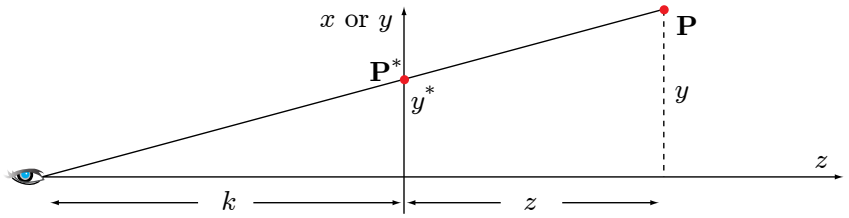


Figure 8.84: Perspective by Similar Triangles.

Step 3. Equation (8.59) can be employed to show that the projection rule of Step 1 results in a projection that satisfies the three conditions above and can therefore be called perspective. Condition 1 says that a distant object should appear small. The object can become distant in three ways:

1. increasing the z coordinates of its points;
2. increasing the x or y coordinates;
3. increasing the value of k.

For large values of z, Equation (8.59) yields small values for x^* and y^*. Specifically

$$\lim_{z\to\infty} x^* = 0 \quad \text{and} \quad \lim_{z\to\infty} y^* = 0.$$

For large values of x or y, imagine two points, $\mathbf{P}_1 = (x_1, y_1, z_1)$ and $\mathbf{P}_2 = (x_2, y_1, z_1)$, on the object that differ only in their x coordinates. They are projected to the two points $\mathbf{P}_1^* = (x_1^*, y_1^*)$ and $\mathbf{P}_2^* = (x_2^*, y_1^*)$, which have identical y coordinates, and the ratio of their x coordinates is

$$\frac{x_1^*}{x_2^*} = \frac{x_1}{(z_1/k)+1} \Big/ \frac{x_2}{(z_1/k)+1} = \frac{x_1}{x_2}. \tag{8.60}$$

Thus, when both x_1 and x_2 grow, the ratio x_1^*/x_2^* approaches 1, which implies that the two projected points $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ get closer. Since $\mathbf{P}_1$ and $\mathbf{P}_2$ are arbitrary points with

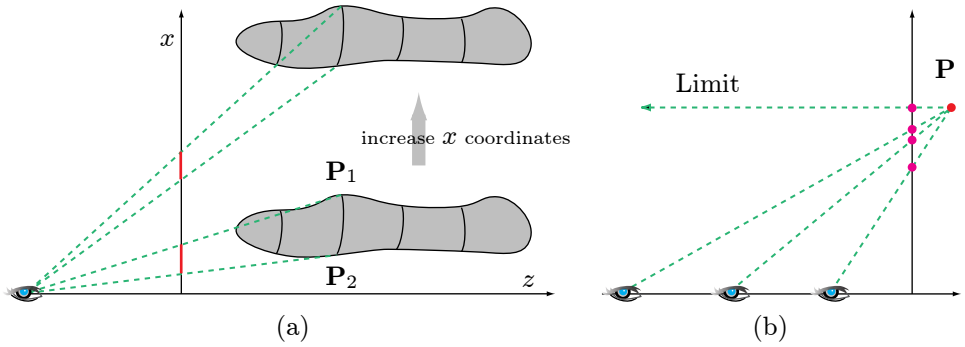


Figure 8.85: (a) Large x Dimensions. (b) Large Values of k.

the same y and z coordinates, this implies that all the points with the same y and z coordinates produce projections that are very close. The entire object seems to have shrunk in the x dimension (Figure 8.85a).

The case where k increases (i.e., the viewer moves away from the projection plane) is different. Figure 8.85b shows how the projection of the object becomes bigger and bigger in this case until, at the limit, when the viewer is at infinity, the projection reaches the actual size of the object. The perspective projection is reduced in this limit to a parallel projection. However, even though the projection itself gets bigger, the viewer sees a small projected object because the projection plane and everything on it look small to a distant viewer.

Condition 2 demands that a distant object feature less perspective. We already know that an object can become distant in three ways each of which is individually treated here.

1. The z coordinates are increased. We select two object points $\mathbf{P}_1 = (x_1, y_1, z_1)$ and $\mathbf{P}_2 = (x_1, y_1, z_2)$ with the same x and y coordinates and different z coordinates. We denote their projected points by $\mathbf{P}_1^* = (x_1^*, y_1^*)$ and $\mathbf{P}_2^* = (x_2^*, y_2^*)$ and compute the ratio x_1^*/x_2^*:

$$\frac{x_1^*}{x_2^*} = \frac{x_1}{(z_1/k)+1} \Big/ \frac{x_1}{(z_2/k)+1} = \frac{z_2+k}{z_1+k}. \tag{8.61}$$

When the z coordinates are increased, this ratio approaches 1, thereby showing that the distance between the projected points is decreased, resulting in less perspective.

2. The x or y coordinates are increased. Equation (8.60) shows that the projected points get closer in this case, too.

3. The value of k is increased. In this case, Equation (8.61) shows that the projected points get closer, again implying less perspective.

Condition 3 is also easy to verify, at least in the case of lines parallel to the z axis. Figure 8.86 shows how a group of lines parallel to the z axis are projected to line segments that converge at the origin.

Step 4. The projection expressed by Equation (8.59) can be included in the general

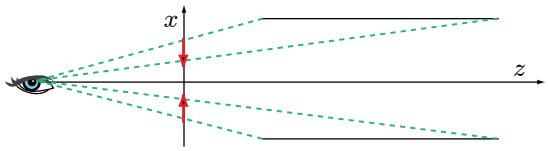


Figure 8.86: Lines Parallel to the z Axis.

4×4 transformation matrix in three dimensions. The result is

$$\mathbf{T}_p = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & r \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.62}$$

A simple test verifies that the product $(x, y, z, 1)\mathbf{T}_p$ yields $(x, y, 0, rz + 1)$ or, after dividing by the fourth coordinate, $(x/(rz+1), y/(rz+1), 0, 1)$. This agrees with Equation (8.59) if we assume that $r = 1/k$. (Recall that k is strictly positive and is never zero because the viewer never presses his eyes to the projection plane.)

It is now clear that there are two more special cases that are geometrically equivalent to our standard position. These are the cases where the viewer is positioned on the negative side of the x axis (or the y axis) at a certain distance from the origin, and the projection plane is the yz (or xz) plane. The object is located on the positive side of the x (or y) axis. These cases correspond to the transformation matrices

$$\mathbf{T}_x = \begin{pmatrix} 0 & 0 & 0 & p \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{T}_y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & q \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where both $1/p$ and $1/q$ are the distances of the viewer from the origin.

Note. The general case, where the viewer can be positioned anywhere and looking in any direction, is covered in [Salomon 11].

We end this discussion of the mathematics of perspective projection with several examples of points projected in the standard position.

▪ Linear example. We arbitrarily select the two points $\mathbf{P}_1 = (2, 3, 1)$ and $\mathbf{P}_2 = (3, -1, 2)$ and the distance $k = 1$. Notice that the z coordinates of these points are nonnegative. The points are projected to

$$\mathbf{P}_1^* = \left[\frac{2}{(1/1)+1}, \frac{3}{(1/1)+1}\right] = (1, 3/2) \text{ and } \mathbf{P}_2^* = \left[\frac{3}{(2/1)+1}, \frac{-1}{(2/1)+1}\right] = (1, -1/3).$$

We now select the midpoint $\mathbf{P}_m = (\mathbf{P}_1 + \mathbf{P}_2)/2 = (5/2, 1, 3/2)$ and project it to

$$\mathbf{P}_m^* = \left[\frac{5/2}{\frac{3/2}{1}+1}, \frac{1}{\frac{3/2}{1}+1}\right] = (1, 2/5).$$

Point $\mathbf{P}_m$ is located on the straight segment connecting $\mathbf{P}_1$ to $\mathbf{P}_2$ (it is the midpoint of the segment) and $\mathbf{P}^*_m$ is on the segment connecting $\mathbf{P}^*_1$ to $\mathbf{P}^*_2$ (although it isn't the midpoint, because it is easy to see that $\mathbf{P}^*_m = 0.4\mathbf{P}^*_1 + 0.6\mathbf{P}^*_2$). The perspective projection of a straight segment is a straight segment, which is why it is done in practice by projecting the two endpoints and connecting them on the projection plane with a straight segment.

▪ Converging lines. We choose an arbitrary point $\mathbf{P}_3 = (0, 2, 3)$ and compute a new point $\mathbf{P}_4 = (1, -2, 4)$ from the relation $\mathbf{P}_4 - \mathbf{P}_3 = \mathbf{P}_2 - \mathbf{P}_1$. The difference of two points is a vector, so this relation guarantees that the vector from $\mathbf{P}_3$ to $\mathbf{P}_4$ equals the vector from $\mathbf{P}_1$ to $\mathbf{P}_2$, or, equivalently, that the two line segments $\mathbf{P}_1\mathbf{P}_2$ and $\mathbf{P}_3\mathbf{P}_4$ are parallel. The two new points are projected to yield

$$\mathbf{P}^*_3 = \left[0, \frac{2}{(3/1)+1}\right] = (0, 1/2) \quad \text{and} \quad \mathbf{P}^*_4 = \left[\frac{1}{(4/1)+1}, \frac{-2}{(4/1)+1}\right] = (1/5, -2/5).$$

The parametric equation of the straight segment connecting $\mathbf{P}^*_3$ to $\mathbf{P}^*_4$ is

$$L_2(w) = w(\mathbf{P}^*_4 - \mathbf{P}^*_3) + \mathbf{P}^*_3 = w(1/5, -9/10) + (0, 1/2), \quad \text{for} \quad 0 \le w \le 1,$$

and the parametric equation of the straight segment connecting $\mathbf{P}^*_1$ to $\mathbf{P}^*_2$ is

$$L_1(u) = u(\mathbf{P}^*_2 - \mathbf{P}^*_1) + \mathbf{P}^*_1 = u(0, -4/3) + (1, 3/2), \quad \text{for} \quad 0 \le u \le 1,$$

the point is that although the original segments $\mathbf{P}_1\mathbf{P}_2$ and $\mathbf{P}_3\mathbf{P}_4$ are parallel, the two projected segments are not parallel. They meet at point $L_1(33/8) = L_2(5) = (1, -4)$.

Another way to prove that the two projected line segments converge is to show that they are not parallel by computing and comparing their directions (or slopes). It's easy to see that $\mathbf{P}^*_2 - \mathbf{P}^*_1 = (0, -4/3)$ but $\mathbf{P}^*_4 - \mathbf{P}^*_3 = (1/5, -9/10)$. Line segment L_1 proceeds straight down, whereas L_2 has a slope of $(-9/10)/(1/5) = -4.5$.

◇ **Exercise 8.42:** Select two line segments that are perpendicular to the line of sight of the viewer, and show that their projections on the xy plane are parallel.

▪ Projecting entire curves. We select the three points $\mathbf{P}_1 = (-1, 0, 1)$, $\mathbf{P}_2 = (0, 1, 2)$, and $\mathbf{P}_3 = (1, 1, 3)$ and compute the Bézier curve $\mathbf{P}(t)$ (Section 8.21) defined by them

$$\mathbf{P}(t) = (1-t)^2(-1, 0, 1) + 2t(1-t)(0, 1, 2) + t^2(1, 1, 3).$$

The midpoint of this curve is

$$\mathbf{P}(0.5) = (-1/4, 0, 1/4) + (0, 1/2, 1) + (1/4, 1/4, 3/4) = (0, 3/4, 2).$$

We now project the three original points and obtain

$$\mathbf{P}^*_1 = \left[\frac{-1}{(1/1)+1}, 0\right] = (-1/2, 0), \quad \mathbf{P}^*_2 = \left[0, \frac{1}{(2/1)+1}\right] = (0, 1/3),$$
$$\mathbf{P}^*_3 = \left[\frac{1}{(3/1)+1}, \frac{1}{(3/1)+1}\right] = (1/4, 1/4).$$

The Bézier curve defined by these points is

$$\mathbf{P}^*(t) = (1-t)^2(-1/2, 0) + 2t(1-t)(0, 1/3) + t^2(1/4, 1/4).$$

The point of this example is that the projection of $\mathbf{P}(0.5)$, which is $(0, 1/4)$, is not located on $\mathbf{P}^*(t)$. This illustrates the nonlinear nature of the Bézier curve (as well as most other curves).

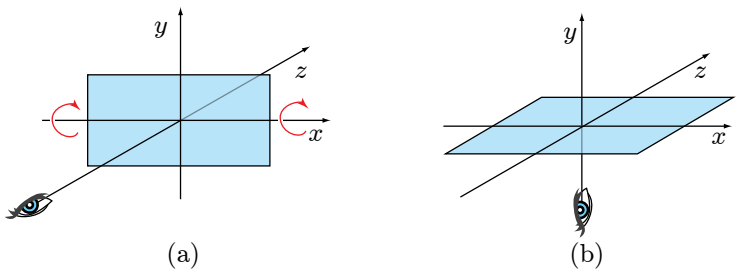
◇ **Exercise 8.43:** Show why point $(0, 1/4)$ is not located on $\mathbf{P}^*(t)$.

■ Transforming and projecting. This example illustrates the advantage of the projection matrix $\mathbf{T}_p$ of Equation (8.62). Given an object, we might want to transform it before we project its points. In such a case, all we have to do is prepare the individual 4×4 transformation matrices, multiply them together in the order of the transformations, and multiply the result by $\mathbf{T}_p$. Assume that we want to apply the following transformations to our object: (1) Rotate it about the x axis by 90° from the direction of positive y to the direction of positive z (Figure 8.87a). (2) Translate it by 3 units in the positive z direction. (3) Scale it by a factor of 1/2 (i.e., shrink it to half its size) in the y dimension. The three transformation matrices are

$$\mathbf{T}_R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{T}_T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 1 \end{bmatrix}, \quad \mathbf{T}_S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and their product with $\mathbf{T}_p$ (we assume $k = 1$, so $r = 1$) produces

$$\mathbf{T} = \mathbf{T}_R\mathbf{T}_T\mathbf{T}_S \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1/2 & 0 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}. \tag{8.63}$$



(a) (b)

Figure 8.87: Rotation about the x Axis.

We can now pick any point on the object, write it as a 4-tuple in homogeneous coordinates, and multiply it by $\mathbf{T}$ to both transform it and obtain its projection. Notice that a point cannot be scaled, but the effect of scaling is to move points such that the entire scaled object would shrink to half its size in the y dimension. As an example, multiplying point $(0,1,-4,1)$ by $\mathbf{T}$ results in $(0,2,0,5)$, which, after dividing by the fourth coordinate, produces the two-dimensional point $(0,2/5)$.

◇ **Exercise 8.44:** Multiply point $(0,1,-4,1)$ by the product $\mathbf{T}_R\mathbf{T}_T\mathbf{T}_S$ and explain the result.

◇ **Exercise 8.45:** The previous paragraph mentions scaling, so let's consider another subtle effect of this simple transformation. The transformation matrix for scaling is

$$\begin{pmatrix} T_1 & 0 & 0 & 0 \\ 0 & T_2 & 0 & 0 \\ 0 & 0 & T_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

When combined with perspective projection, it yields

$$\begin{pmatrix} T_1 & 0 & 0 & 0 \\ 0 & T_2 & 0 & 0 \\ 0 & 0 & T_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & r \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} T_1 & 0 & 0 & 0 \\ 0 & T_2 & 0 & 0 \\ 0 & 0 & 0 & T_3 r \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hence, a point $(x,y,z,1)$ is transformed to $(T_1x, T_2y, 0, T_3rz+1)$, which implies

$$x^* = \frac{T_1 x}{T_3 rz + 1}, \qquad y^* = \frac{T_2 y}{T_3 rz + 1}.$$

In the special case of uniform scaling, $T_1 = T_2 = T_3 = T$, we get $x^* = x/(rz + 1/T)$, $y^* = y/(rz + 1/T)$. The problem is that when T gets large (large magnification), $1/T$ becomes small, resulting in

$$x^* \approx \frac{x}{rz} = \frac{xk}{z}, \qquad y^* \approx \frac{y}{rz} = \frac{yk}{z}.$$

We don't seem to get the expected magnification. What is the explanation?

The rightmost column of matrix $\mathbf{T}$ of Equation (8.63) is important and is crucial in any discussion of the general perspective projection (as e.g., in reference [Salomon 11]). The three top elements of this column are 0, 1, and 0. In the general transformation matrix, these elements are denoted by p, q, and r. Thus, element q of matrix $\mathbf{T}$ is nonzero. It has already been mentioned that element r of matrix $\mathbf{T}_p$ is nonzero because the viewer is positioned on the z axis. The reason that element q of matrix $\mathbf{T}$ is nonzero is the rotation about the x axis. We can interpret this rotation either as a rotation of the point or as a rotation of the coordinate system. In the latter case, this rotation has changed the projection plane from the xy plane to the xz plane and has also moved

the viewer (because the viewer and the projection plane should be considered a single rigid unit) from his standard position on the z axis to a new location on the y axis (Figure 8.87b). The fact that q is nonzero implies that the y axis now intercepts the projection plane.

◇ **Exercise 8.46:** Compute the coordinates of the object point $\mathbf{P}$ that happens to be projected to the origin after the three transformations.

■ Negative z coordinates. It has already been mentioned several times that the viewer and the object have to be located on different sides of the projection plane. In the standard position, this means that all the object points must have nonnegative z coordinates. This example shows what happens when object points have invalid coordinates. Figure 8.88a shows the two points $\mathbf{P}_1 = (0, 1, -1)$ and $\mathbf{P}_2 = (0, 1, 1)$ and a viewer located at $(0, 0, -3)$. When Equation (8.59) is used to project the two points, the results are

$$\mathbf{P}_1^* = \left[0, \frac{1}{(-1/3)+1}, 0\right] = (0, 3/2, 0) \quad \text{and} \quad \mathbf{P}_2^* = \left[0, \frac{1}{(1/3)+1}, 0\right] = (0, 3/4, 0).$$

The result seems to make sense, but Figure 8.88b shows that when $\mathbf{P}_1$ is moved to the left (i.e., toward larger negative z values), its projection climbs up the y axis quickly and without limit, thereby creating a distorted projection of the entire object. When $\mathbf{P}_1$ is located right over the viewer [when it is moved to $(0, 1, -3)$], its projection is undefined, and when it is moved farther to the left, its projection becomes negative. In such a case, those parts of the object that are in front of the viewer are projected right-side up but distorted, and those parts that are behind the viewer are projected upside down.

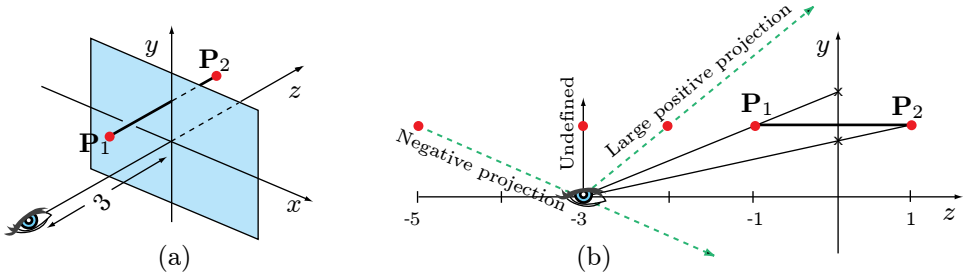


Figure 8.88: Perspective Projection with Negative z Coordinates.

8.26 Slanted Squares with Integer Corners

Geometry leads to number theory. Given the dotted array of Figure 8.89 where the dots have integer coordinates, it is easy to see that we can draw many squares on it whose sizes are integers and whose areas are the square numbers n^2 for integer n values. However, in addition to the obvious, "straight" squares, it is also possible to draw slanted squares whose corners are located on the dots, and the topic of this discussion is to find a general expression for the areas of all the possible squares on such a grid, and to determine what areas are impossible for slanted squares. This geometry problem is solved by results from number theory, further demonstrating the tight relationship between seemingly remote areas of mathematics.

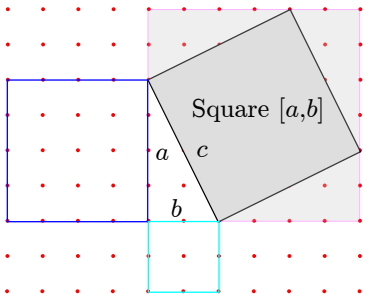


Figure 8.89: Slanted Squares on a Grid.

The gray square of Figure 8.89 Illustrates the problem. We start at a dot, move a steps down and b steps to the right, arriving at another dot. Slanted segment c becomes one side of a square that we can refer to as an $[a, b]$ square (in the figure, a $[4, 2]$ square that is embedded in a larger 6×6 square). The area of the $[4, 2]$ square is therefore the area 6×6 of the large square, minus the areas of the four triangles, which are $(4 \times 2)/2$ each.

It is trivial to extend this example to the general case of a slanted $[a, b]$ square. Such a square is embedded in a larger $(a + b)(a + b)$ square, surrounded by four right triangles of sides a and b each. The total area of the triangles is $4ab/2 = 2ab$, which gives the slanted $[a, b]$ an area $(a + b)(a + b) - 2ab = a^2 + b^2$. One special case is when either a or b are zero, where the square turns out to be straight. Another special case is $a = b = 0$, where the square is reduced to a single point and its area is zero.

As an aside, Figure 8.89 also illustrates the relation $c^2 = a^2 + b^2$, thereby providing a (geometrical) proof of Pythagoras theorem.

It is simple to check manually which integers are not of the form $a^2 + b^2$, and Figure 8.90 lists a few. However, a general rule is needed, and the rule that is mentioned in reference [ImpossibleSquares 20] is a known result from number theory. It states that if an integer n has a prime factor of the form $(4k + 3)^p$ for an odd p, then n is not of the form $a^2 + b^2$ for integers a and b.

For example, the prime factors of 21 are 7 and 3, and since 7 equals $(4 \times 1 + 3)^1$, we conclude that 21 is not of the form $a^2 + b^2$ and is in fact on the list of Figure 8.90.

```
l={};
Do[c=0;
Do[ If[a^2+b^2==n,c=1],{a,0,n-1},{b,0,n-1}];
 If[c==0,AppendTo[l,n]]
 ,{n,1,25}]
l

{1,3,6,7,11,12,14,15,19,21,22,23,24}
```

Figure 8.90: Integers Not of the Form $a^2 + b^2$.

On the other hand, the prime 13 is of the form $2^2 + 3^2$, which is why a slanted square of area 13 exists.

8.27 Area of regular polygons

Figure 8.91 shows that in a regular polygon $\tan \pi/n = (l/2)/a = l/(2a)$, which is why the apothem a equals $l/(2\tan(\pi/n))$. (An irregular polygon does not have an apothem.) The area of the gray triangle is $(a \cdot (l/2))/2 = l^2/(8\tan(\pi/n))$, so the area of the entire polygon, which consists of $2n$ such triangles, is

$$\frac{n \cdot l^2}{4\tan(\pi/n)}.$$

Notice that as the number of sides n grows, the polygon approaches a circle. Figure 8.91 also shows that in a circle, the radius R equals the apothem a. The area of a circle is πR^2, so we divide the area of a polygon by a^2 to obtain

$$\frac{n \cdot l^2}{4\tan(\pi/n)}/a^2 = n\tan(\pi/n),$$

and use mathematical software to verify that this expression approaches π for large n.

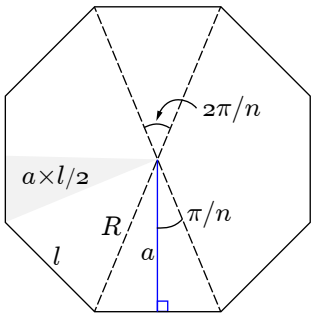


Figure 8.91: A Regular Polygon.

8.28 The Fourth Side of a Triangle?

A triangle has many centers. Its orthocenter is the point where the three altitudes (the straight segments that are perpendicular to the sides of the triangle) intersect. These three points, together with the three vertices of the triangle and the three midpoints of its edges, are located on the circumference of a circle, the so-called nine-point triangle circle.

(Given a triangle, its nine-point circle is the circle that passes through the midpoints of the triangle's sides, the feet of its altitudes, and the midpoints of the line segments connecting each vertex to the orthocenter. A total of nine significant points.)

Now imagine the nine-point circle, the incircle (which is inscribed in the triangle), and the outer, or circumcircle (that circumscribes the triangle). Each side of the triangle has (1) its endpoints on the circumcircle, (2) its midpoint on the nine-point circle, and (3) is tangent to the incircle (Figure 8.92a).

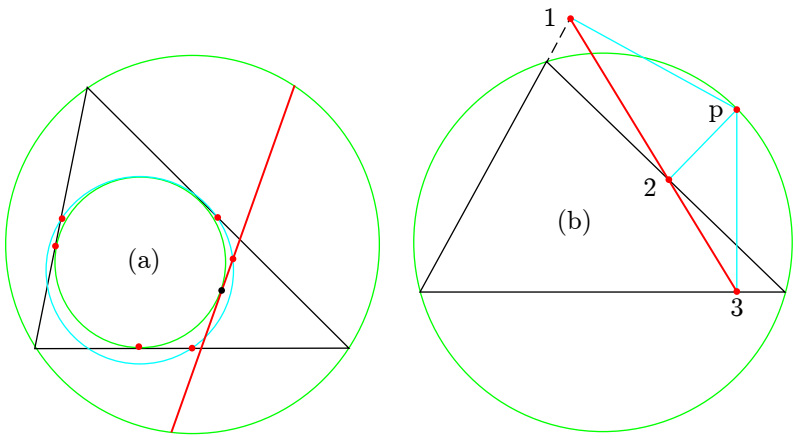


Figure 8.92: (a) Sherman Fourth Line. (b) Simson Triangle.

The surprise is that there is a fourth line that satisfies the three conditions above. This line segment, red in the Figure, was discovered by B. F. Sherman in 1993 (references [Sherman 93] and [Yiu 12]). It is termed the pedal line of the triangle, but can, with some justification, be called the fourth side of the triangle.

In part (b) of the figure a point `p` is chosen on its circumcircle. We notice that points 1, 2, and 3, which are located on the sides of the triangle and are closest to `p`, are collinear (the red line in the figure). This is the Simson line of the triangle.

I think I still like science and art better, but geometry is a big improvement over algebra.

—Laurie Hernandez.

9
Puzzles

Every culture has its own myths, but it is known that certain mythological elements and motifs recur in many cultures; they are universal. The modern theory of universal myth was founded and developed by many researchers, but its most well-known proponent is Joseph Campbell [Campbell, Joseph 08].

> Mr. Casaubon, as might be expected, spent a great deal of his time at the Grange in these weeks, and the hindrance which courtship occasioned to the progress of his great work—the Key to all Mythologies—naturally made him look forward the more eagerly to the happy termination of courtship.
>
> —George Eliot, *Middlemarch* (1871–72).

Among the universal myths are stories of the creation of the world, flood, drought, and famine, seasonal death and rebirth, the origin of fire, the relations between the dead and the living, and of special interest to us, puzzles. Puzzles are found in the stories and legends of many cultures throughout history. It seems that the human brain is attracted to puzzles, riddles whose solutions require much thinking and deep concentration, but that can nevertheless be solved by anyone with average intelligence and with no specialized knowledge.

9.1 Examples of Puzzles

Today, the English word *puzzle* often refers to jigsaw puzzles, but there are many other types of puzzles, such as picture puzzles, logic puzzles, chess problems, trivia, anagrams, and riddles. This chapter is concerned with mathematical puzzles, of which currently the most popular is Sudoku.

9.1.1 Dissecting a square

It is easy to partition or dissect a square in many ways. Figure 9.1a shows how a square can be cut into distinct rectangles. Part (b) of the figure shows how easy it is to partition a square into smaller squares many of which are identical. What about a dissection that partitions a square into several distinct squares?

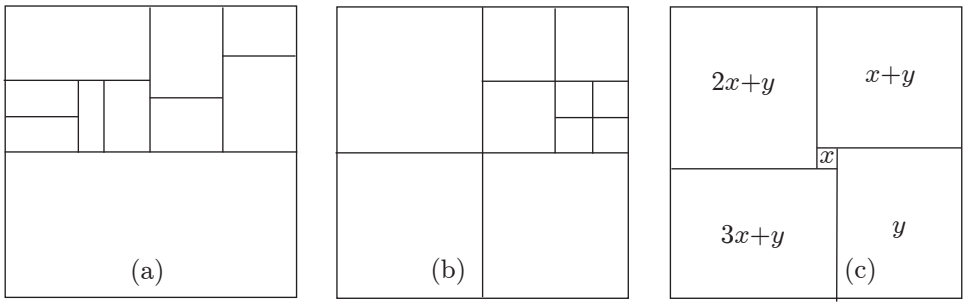


Figure 9.1: Dissecting a Square.

This is the kind of problem that fires the imagination of mathematicians. The first step in solving it is to define rigorous terms. A dissection that cuts a square into distinct squares is called perfect. If no subset of the squares forms a rectangle, the dissection is also referred to as simple. How can we find a prefect simple dissection of the square? Figure 9.1c illustrates an approach, and also shows why this is not a trivial problem.

We tentatively try to create a simple perfect 5-square dissection. We place a square of side y at the bottom-right corner and another square, of side x, to its left. This implies that the square at the top-right corner has a side of $x + y$ units, the square at the top-left has a side of $2x + y$ units, and the last square, on the bottom-left corner, has a side of $3x + y$. Everything seems fine until we try to draw the entire square. It is easy to see that the top and bottom of the large square are each $3x + 2y$ units long. The left edge, however, is $5x + 2y$ and the right edge is $x + 2y$ units. The only solution of the equation $5x + 2y = x + 2y$ is $x = 0$. Thus, such a square cannot be constructed.

The problem is not trivial, but many solutions were found over years of patient search. The smallest one starts with a 112×112 square which is dissected into 21 distinct squares (Figure 9.2, where the numbers are the sides of the squares). This solution is due to [Duijvestijn 78] and it seems to be the smallest one possible. Bigger squares can yield simple perfect dissections in several ways. The number of such dissections for 21, 22,... dissections is 1, 8, 12, 26, 160, 441,... (this is sequence A006983 in [Sloane 08]).

The square of Figure 9.2 features complex structure and looks strange and asymmetric in spite of its title as simple and perfect. However, there is a simple notation for describing the structure of such a nested square. We look for a group of adjacent squares that are aligned at the tops and write their sides in square brackets. Thus, [50, 35, 27], [8, 19], [15, 17, 11], [29, 25, 9], [7, 18], [16], [42], [4, 37], and [33]. Notice how the groups are arranged from top to bottom, starting with [50, 35, 27] and ending with [33]. If two groups are at the same height, they are arranged left to right.

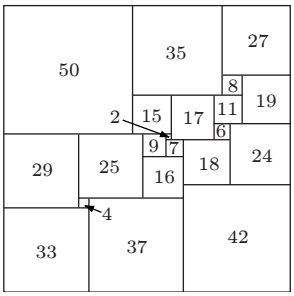


Figure 9.2: Minimal Simple Perfect Dissection.

9.1.2 The 15 Puzzle



The image shows the 14-15 puzzle, Sam Lloyd's version of the well-known 15-puzzle. Fifteen square tiles, numbered 1 to 15, are arranged in a 4×4 frame that has room for 16 tiles. Thus, there is a hole left into which a tile can be slid. The tiles are arranged by numbers top to bottom the left to right, except that 14 and 15 are swapped. The goal of this puzzle is to slide tiles until all 15 tiles are in perfect sequence, or equivalently, to somehow swap 14 and 15. This problem is unsolvable, which is proved by considering permutations and their parities. An equivalent version is to start with a perfectly arranged tile and ask if there is a way to swap 14 and 15. A more general question is whether it is possible to get every possible pattern of tiles from the original pattern.

Sam Lloyd himself offered $1000 to anyone who could solve his version, but it had been known even before his time (by Johnson and Story, in 1879, see [Johnson.Story 79]) that this is impossible.

To understand the problem, we think of the empty space as occupied by an imaginary tile, and then realize that any tile movement, in any direction, amounts to swapping the imaginary tile with one of its neighbors. Imagine a number of such swaps where the imaginary tile ends up in its original place at the bottom-right corner. A little thinking should convince the reader that such a process requires an even number of swaps. This is because each swap that moves the imaginary tile to the left, must have a corresponding swap that moves it to the right, and the same is true for vertical swaps.

However, it can be proved, using permutations, that any pattern that can be achieved with an even number of swaps can only be achieved with an even number of swaps, and similarly, any pattern that can be achieved with an odd number of swaps can only be achieved with an odd number of swaps. This partitions the space of possible patterns into two parity classes, even patterns and odd ones. The 15 puzzle is limited to even number of swaps, and swapping 14 and 15 is one by a single swap, which cannot be performed in the 4×4 frame of the puzzle.

9.1.3 The 16-room puzzle

Figure 9.3 shows a simple 4×4 plan with 16 squares or rooms, numbered and fully

connected with passages. The goal of the puzzle is to figure how the occupant of room 1 can move from room to room, visiting every other occupant exactly once, and finally leaving the maze through room 16.

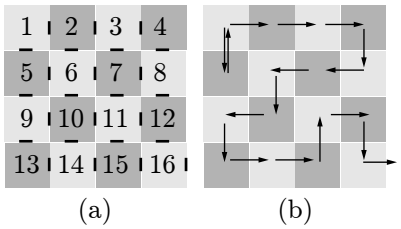


Figure 9.3: The 16 Room Maze.

This is a simple puzzle that illustrates the meaning of the phrase "thinking outside the box. The point is that when solving a puzzle that is based on a diagram, we tend to find solutions that are confined to the diagram, whereas many solutions require moving outside the area occupied the diagram.

The simple solution of this puzzle is shown in part (b) of the figure. The occupant of room 1 should move to room 5 (or, alternatively, to room 2), then come back to his own room, go to room 2 (or room 5), and proceed as in the figure.

◇ **Exercise 9.1:** Given the two circles and line segment of Figure 9.4, your task is to draw a curve tracing them without lifting your pen, and without retracing or crossing anything you have already drawn.

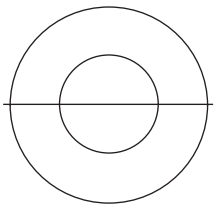


Figure 9.4: Figure for Exercise 9.1.

9.1.4 Vanishing Area Puzzle

Figure 9.5 shows an 8×8 square dissected into four pieces that are then assembled into a 5×13 rectangle. This simple puzzle, due to Sam Lloyd, seems magical because the area of the original square is 64 units while the rectangle's area is 65 units. The figure also shows an alternative way to assemble the four pieces into a $5 \times 6 + 3 + 5 \times 6$ figure whose area is 63.

It turns out that this is only an optical illusion. The assembled pieces do not fit precisely together, a fact which is concealed by the thick lines that are noticeable in the figure.

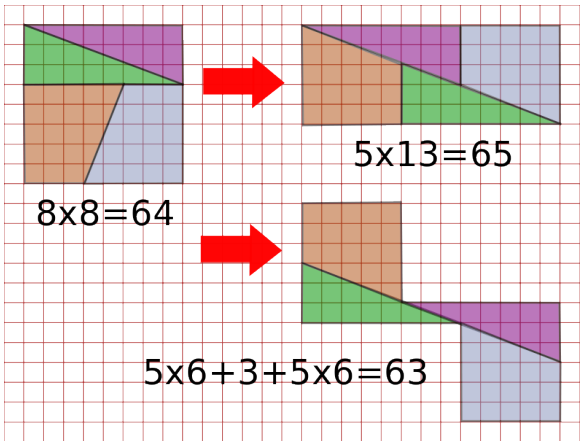
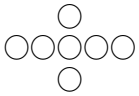


Figure 9.5: Sam LLoyd Dissection.
Courtesy of Wikipedia user Yoni Toker
`https://en.wikipedia.org/wiki/File:Loyd64-65-dis_b.svg`

This puzzle becomes much more "palatable" when demonstrated by squares of chocolate.

◇ **Exercise 9.2:** Figure 9.6(left) is a triangle consisting of six geometric objects and two holes, for a total of 60 chocolate squares. The right part of the figure suggests that the six objects, which previously covered 58 of the 60 squares, can now cover the entire area of 60 squares. Explain this puzzle.

◇ **Exercise 9.3:** This is a traditional coin puzzle which can easily be solved by "thinking outside the box." Given seven coins arranged as in the figure, move two coins to end up with two rows of five coins each, one vertical and the other horizontal, crossing at their centers.



9.1.5 The Euler Trail puzzle

Given the 16-segment boxlike structure of Figure 9.7(a), draw a continuous curve that crosses each of the 16 boundary segments exactly once, without lifting the pen. The green curve shows a wrong solution. What is your solution? This is similar to the Königsberg bridge problem, solved by Euler in 1735.

The solution is simple; there is no solution. Figure 9.7(b) shows all the possible ways to connect the 16 segments (in red) that pass through the 16 boundaries. The connecting segments meet at five points (red circles) and the important fact is that three of those points have an odd number (five) of red segments. As an example, following red segment 1 as it enter the structure, we arrive at a circle where red segments 2, 3, 4, and 5 meet.

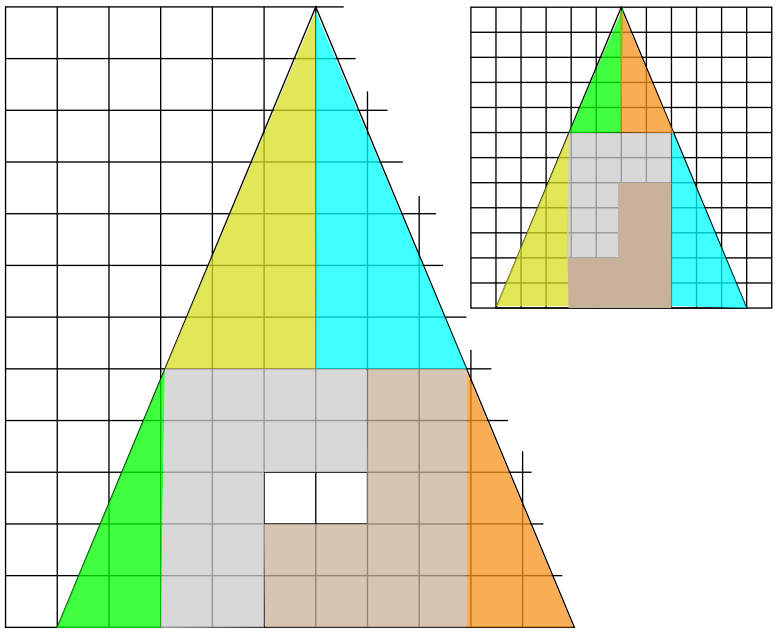


Figure 9.6: 60 Chocolate Squares.

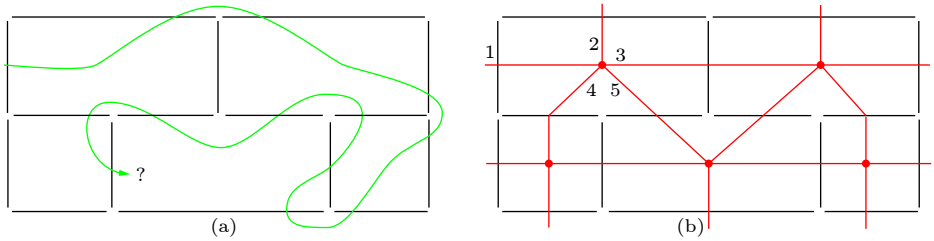


Figure 9.7: Euler Trail Diagram.

If a solution exists, it must satisfy the following: When a red segment enters a circle, it must leave it, and it must do so through another "fresh" red segment, one that hasn't yet been traversed. Therefore, the number of red segments that meet at any red circle must be even. The only exceptions are the red circle that is the start of the traversing curve and the last circle in the traversal. There can only be two such circles, but our structure has three, which proves that there is no solution to this simple problem.

Note. When a teacher showed us this puzzle in elementary school, she claimed that Einstein was the only known person to solve it on his first try. Even the great Euler had to spend some time on it. Now that you know how simple the solution is, you may doubt this claim.

> Do not worry about your difficulties in Mathematics.
> I can assure you mine are still greater.
> —Albert Einstein.

9.1.6 A Submarine Hunt

The 2009 best seller novel (and a later movie) *The Hunt for Red October*, by Tom Clancy, is about the American CIA hunting a Soviet submarine in the Atlantic. The puzzle presented here is also about hunting a submarine, but is very different in the following respects: (1) It takes place on the one-dimensional integer number line and (2) It is infinite.

Figure 9.8 illustrates a submarine starting at an integer position x_0 on the infinite integer number line, and moving at a constant speed v (also an integer) from integer to integer on this line as time proceeds in fixed steps. Being a submarine, it is invisible, but we have a detector that can check an arbitrary integer location on the line in each time step and tell whether the submarine is there. Both the start position and the speed of the submarine are unknown, which means that at any time it can be anywhere on the line. Note that the speed can be negative.

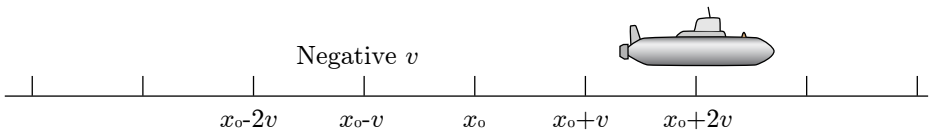


Figure 9.8: Submarine Time Steps.

Our job is to find an algorithm that will always locate the submarine in a finite number of steps. Like many problems that initially seem impossible, this one has a simple, surprising solution that is based on the zigzag scanning technique, covered on Pages 245 and 253.

The problem is two-dimensional in the sense that it depends on the speed v and the initial position x_0. Once these are known, the position x of the submarine at any time t is given by $x = x_0 + vt$. Thus, each possible submarine path is associated with a pair (x_0, v) of initial location x_0 and a certain constant speed v. We therefore start by looking at the two-dimensional space whose coordinates are x_0 values (the rows) and v values (the columns). A point in this space has integer coordinates (x_0, v), but notice that each coordinate is signed and can be negative. The space shown in Figure 9.9 has the location and speed values interleaved in the form 0, 1, -1, 2, -2, 3, ... This causes the space to be infinite only to the right and down. Point $(0, 0)$ is located at the top-left corner.

The points are now connected and numbered in a zigzag pattern, with the top-left point numbered 0. This should be compared with the similar pattern illustrated in figures 4.1 and 4.5.

The search for the submarine now proceeds in time steps, where in step n we visit point n on the zigzag path in the figure. The (x, y) coordinates of this point become a pair (x_0, v) that is associated with a particular submarine path, and we check for the

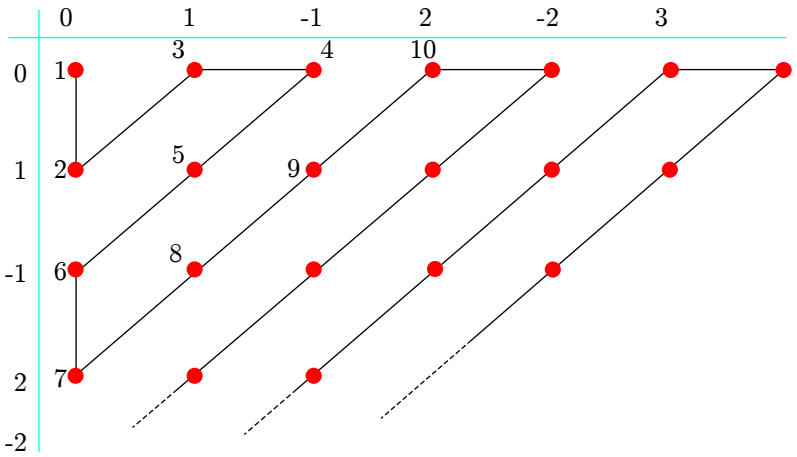


Figure 9.9: A Zigzag Search Pattern.

presence of the submarine at this pair (i.e., a submarine that started at that x_0, travels at speed v, and is now at $x_0 + vn$). If the submarine is not found, we know that it will not be found in this particular pair in future time steps. This is because each possible submarine path is associated with a pair (x_0, v) and, at time t_n, the submarine must therefore be at location $x_0 + vn$. If a submarine is not found at this location, pair (x_0, v) can be ignored.

Thus, for example, in step t_9 we visit the pair $(1, -1)$ and should therefore check location $1 - 1 \cdot 9 = -8$. If the submarine isn't at location -8 on the number line, then it is not associated with point $(1, -1)$ and we proceed to step t_{10}, where we visit pair $(0, 2)$, and so have to check location $0 + 2 \cdot 10 = 20$ on the number line. Continuing in this way we will never visit the entire zigzag path, but given enough time, we will find the pair (x_0, v) that is associated with our submarine.

◇ **Exercise 9.4:** Our space is two-dimensional and is infinite in both dimensions. The zigzag search pattern places the center of the space at the top-left corner of the search path. Show how it is possible to search such an infinite space even if the center of the space is placed at the center of the search path, and all four sides of the search diagram are infinite.

9.1.7 Where is the father?

A mother M is 21 years older than her child C. Six years from now, her age will be five times her child's age. The question is Where is the father F right now?

This is a weird puzzle. There seems to be no relation between the ages of M and C and the current location of F, but let's start with what data we have. It's easy to set and solve the two age-related equations $M = C + 21$ and $M + 6 = 5 \times (C + 6)$. The solutions are $M = 79/4$ and $C = -3/4$, and what should immediately catch our eye is the negative age of the child. A negative value means that the child hasn't been born yet, and the precise value is $3/4$ of a year, or nine months. The surprising, unexpected

conclusion is that the child is being conceived right now, which means that the father F must be *with* the mother M. Puzzle solved and let's give them the privacy they deserve.

9.1.8 She plays the violin?

It is the height of summer, and Violetta, a night receptionist at the Imperial hotel in Sacramento, is happy to meet Rodolfo, a new hotel guest. His name implies that he is Italian, like Violetta's parents, and he is quick to confirm this. During their conversazione, they find out that each has three children. "My figli (sons) are 5, 7, and 9," says Rodolfo. "How old are yours?" But Violetta, a mathematician during the day, answers with a riddle. "The product of my figlie (daughters') ages is 36, and their sum equals your room number."

"I know my room number," says Rodolfo, "so give me a minute." However, after half an hour, he comes back to the reception desk, puzzled. "The information you gave me isn't enough," he says. "How about another hint?" "Con piacere (with pleasure). The hint is: My eldest daughter plays the violin." "Great," says Rodolfo. "Now I know their ages, and I can sleep well tonight. Does the air conditioning work?"

The obvious questions are: (1) What are their ages, and (2) how did the hint help Rodolfo?



Figure 9.10: Image by MS Copilot, Aug. 2025.

The violin hint is puzzling (would a trombone have made a difference?) but let's start with what we have. Table 9.11 lists all the ways in which three positive integers can produce 36 when multiplied. Their sums S are also listed.

The next hint is the room number. The table shows that the possible room numbers are 10, 11, 13, 14, 16, 21, and 38. We don't know Rodolfo's room number, but why can't he solve the problem even though he knows the number? How does the violin hint help him?

By now, some anxious readers may have seen the solution. The room number is 13, but this number corresponds to two possible sets of ages, 2, 2, 9, and 1, 6, 6, which is

		P	S			P	S
1	1	36	38	2	2	9	13
1	2	18	21	2	3	6	11
1	3	12	16	3	3	4	10
1	4	9	14	1	6	6	13

Table 9.11: Products and Sums of Three Ages.

why Rodolfo needed more information. The keyword in the hint was not "violin," but "eldest." In English, the word "eldest" refers to one out of three or more persons. The set 1, 6, 6 of ages has two eldest daughters, so the only choice is 2, 2, 9. Clever!

The following puzzles were mostly drawn from references:

Gratzer, George A. (2011) *Train Your Brain: A Year's Worth of Puzzles*, CRC Press.

Friedland, Aaron J. (1970) *Puzzles in Math & Logic*, Dover publications.

Brecher, Erwin (1996) *Puzzles, Mathematical Diversions and Brainteasers*, St. Martin Griffin, New York.

King, Lloyd (1996) *Puzzles for the High IQ*, Sterling, NY.

Mathematicians' ages. Three mathematicians, X, Y, and Z, are trying to tell us their ages, but they use different number bases. Here is what they say, and keep in mind that the number bases as well as their ages, are integers.

X declares: "My age depends on a one-digit integer x in number base q, and on the three one-digit numbers $b = x + 2$, $c = x + 4$, and $d = x + 6$. The age is both the two-digit number xb and the product $c \cdot d$."

Y says: "My age similarly depends on a one-digit integer x, but in a different number base, and on the three one-digit numbers $b = x + 1$, $c = x + 2$, and $d = x + 3$. The age is both the two-digit number xb and the product $c \cdot d$."

Z states: "My age conforms to the guidelines of X, and I am the youngest of the three. That's all I will say."

Once you have figured out their ages, the next question is how to locate their precise birthdays and addresses, so you could sent them birthday cards.

Solution. We follow X's information to write

$$qx + (x + 2) = (x + 4)(x + 6), \tag{9.1}$$

$$x(q + 1 - x - 10) = 22. \tag{9.2}$$

The right-hand side of Equation (9.2) is 22, an integer, which implies that its left-hand side is also an integer. We can therefore write this equation in the form $x \cdot y = 22$, which makes it obvious that x is a factor of 22 and is therefore one of 1, 2, 11, and 22. We check each of those possibilities in turn.

If $x = 1$, then $q + 1 - x - 10 = 22$, implying that $q = 32$.

If $x = 2$, then $2q + 2 - 4 - 20 = 22$, implying that $q = 22$.

If x is 11 or 22, then Equation (9.1) implies that q is too big to be someone's age.

Thus, two possibilities remain. The first one, ($x = 1$, $q = 32$), produces

$$32 \cdot 1 + (1 + 2) = (1 + 4)(1 + 6) = 35 \text{ years},$$

while the second one ($x = 2$, $q = 22$), yields

$$22 \cdot 2 + (2 + 2) = (2 + 4)(2 + 6) = 48 \text{ years}.$$

The obvious conclusion is that X is 48 years old in base q (24 decimal) and Z, who is younger, is 35 years old in base q (13 decimal).

The case of Y is slightly different. Assuming that he used number base p, we can write $p \cdot x + (x + 1) = (x + 2)(x + 3)$ or $x(p + 1 - 5 - x) = 5$. It is clear that x must be a factor of 5, which makes it either 1 or 5. In the former case, $p = 10$, corresponding to a decimal age of $3 \cdot 4 = 12$ years. This result must be rejected, since Z said that he is the youngest. The latter case starts with the same $p = 10$, which yields age $7 \cdot 8 = 56$ years.

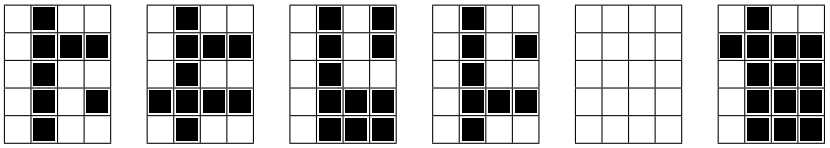
Rope around the equator. Imagine a thin string placed tightly around the equator, at a height of one meter. This requires strong, high-precision supports. The supports are now removed and the string is stretched and placed on the ground around the equator. How much loose string is left in the end?

Solution. Assuming that the radius of the earth is R meters, the length of the original string is $2\pi(R+1)$, while its final length is the circumference of the earth, $2\pi R$. The difference is $2\pi \approx 6.28$ meters. Notice that this result is independent of R. It is true for strings placed around the sun, the solar system, and even around a single atom. A surprising result obtained by elementary mathematics.

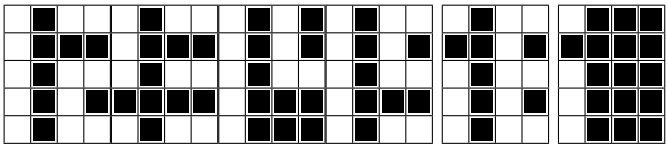
Mathematicians with hats. My three mathematician friends were bored, hence this puzzle. I had three white hats and two black hats. I arranged my friends in front of me in a straight line, one behind the other, so they could only see straight ahead. Each could see only the one or two persons in front of him. I then placed hats on them in order from back to front, so each could see just the hats in front of him. "Now," I said, "can you tell me your hat color?" The person in the rear immediately said "I cannot." Then the one in the middle admitted the same. Finally, the person in front, who could not see any of the hats, told the correct color of his hat. Explain how he knew that.

Solution. The person sitting furthest back could see two hats, yet he couldn't figure out his hat color. He must therefore have seen either two white hats or one white and one black. Upon hearing this, the person in the middle realized that he and the one in front of him must be wearing either two white hats or one white and one black. The fact that he could not decide on his hat color implied that the front person was wearing a white hat. Thus, the person in front was able to deduce that his hat was white.

Black or white? Which cells in the fifth block should be black?



Solution. Once we concatenate the six blocks, it becomes clear that the white cells are supposed to represent the digits 1 through 6. The fifth block is slightly separated below to show its content.



Trisecting an angle. Rodolfo, a student, claims to his professor, "I've just constructed these two angles α and β and I have a proof that β is exactly $1/3$ of α. Moreover, I did this with just a ruler and compass, and angle α is arbitrary." The professor smiles and says, "you should know better. Already in 1837, Pierre Wantzel proved that angle trisection with just a ruler and compass is possible for only certain angles and not in general." "But I did it," repeats Rodolfo.

Solution. How did Rodolfo do it? Easy. He started with an arbitrary angle β and tripled it to get α.

Sum of primes. Alfredo has recently confessed to me that on his first date with Violetta he talked to her about his cousins. "You may not believe it," he said to her, "but my age and the ages of my three cousins are all prime numbers and add up to 50." To his surprise, instead of slapping him and running away, she immediately countered with "I know your age and I know about prime numbers, so I have easily figured out their ages. Now let's talk about our future together." Alfredo was surprised by her quick and intelligent response, but I have commented that she could have figured out their ages even without knowing his age. How is this possible?

Solution. We assume that the ages are integers and none is 1 year. The relevant primes (all under 50) are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, and 47. The only possible combination is for Alfredo to be 43 and for his cousins to be 2, 2, and 3 years old. Any other prime value of Alfredo's age would be associated with several sets of his cousins' ages. For example, if Alfredo's age is 11, then his cousins could be either 3, 5, and 31 or 3, 7, and 29.

Capitals. In what 6-letter country can the following capitals be found? Paris, Oslo, Lima, Ankara, Nairobi, Dublin?

Solution. In Poland. The capitals in the puzzle refer to capital letters, not capital cities.

Clock hands overlap. This puzzle is a darling of recruiting officials in many companies. These clever individuals try to pose questions that reveal the intelligence of job candidates, rather than their comprehensive knowledge or computational prowess.

Imagine an analog clock with just two hands, hours and minutes. At twelve o'clock, the two hands overlap. The next time they overlap is a little after 1 o'clock, but exactly when?

Solution. To find out, we first consider the number of times they overlap during 12 hours. The first overlap is at 12. The second one is a little after 1. The third one is

a (longer) little after 2, and so on, until the 11th overlap, which occurs a long time after 10 o'clock.

A total of eleven overlaps in twelve hours implies a time between consecutive overlays of $12/11 \approx 1.0909$ hours or $1.0909 \times 3600 \approx 3927.27$ seconds. Thus, the times of overlap are 12, 1:05.27, 2:10.55, 3:16.22, 4:21.49, 5:27.17, 6:32.44, 7:38.11, 8:43.38, 9:49.06, and 10:54.33.

Guilty and free. A man is accused of first-degree murder. The judge, after examining all the evidence against the accused, agrees. "There is no doubt," he says, "that this is one of the worst murders in our country's history." "In addition, there are no mitigating circumstances, so I am forced to impose the maximum sentence. However, there is a reason, unrelated to the murder, because of which I have to let the accused go free, although not acquitted." What was the reason?

Solution. The accused was half of a Siamese twins.

Cutting circles. Show how to cut the five circles of Figure 9.12a into two equal-area parts with one line. Do the same for part (b) of the figure (this is easy).

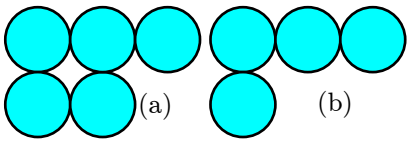


Figure 9.12: Partition Circles.

Solution. Add a sixth circle as in Figure 9.13, and draw a line through the centers of circles 1 and 6. This partitions the six circles into two equal parts. Notice that this line also cuts circle 6 in two equal parts, so just remove it. The solution for part (b) is trivial.

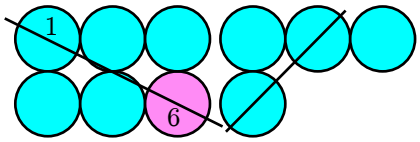


Figure 9.13: Partition Circles.

The seventh. What is the missing digits in the sequence `014285?14`?

Solution. The digit is 7, because $1/7 = 0.14285714\ldots$.

Perfect generosity. At the recent campaign for the eradication of boils, The well-known philanthropist Philo Trope announced that he would match the total of all the other contributions. Upon hearing this, his friend Archie Nemy has fainted. Why?

Solution. Mr. Nemy had made the same pledge. Now, when someone else donates any amount, however small, both Trope and Nemy would have to donate their entire fortunes, each trying to match the other's donation.

Perfect weather. Reporter: What were the noon temperatures in the last five days?

Meteorologist: You reporters ask too many questions, so let me tell you this. The temperatures were distinct, they were whole numbers, and their product was 12. Your task (you are not a reporter, are you?) is to figure out the individual temperatures.

Solution. This is confusing, since the product of even the five smallest, distinct integers is greater than 12. Thus, out-of-the-box thinking is required, which should give the solver the idea that temperatures can be negative. Once this is grasped, we can easily come up with -2, -1, 1, 2, and 3 as the solution. In fact, any positive integer that can be expressed as a product of five distinct integers can be a candidate instead of 12. Even a number such as $18 = 2 \cdot 3 \cdot 3$ can be chosen, with temperatures -1, 1, 2, 3, and 3.

They live where? Erica lives in Washington and Tina lives in Buenos Aires. Where does Mark live?

Solution. He lives in Copenhagen. Each of the names above is the tail-end of the name of the country in whose capital city they live.

A mystery sequence. Each digit in the sequence `123 451 46733 8197 80?3` corresponds to a letter. Which letter should be at the "`?`"

Solution. The sequence of letters and their corresponding digits is shown here:

```
ONE TWO THREE FOUR FIVE
123 451 46733 8197 80?3
```

Hence, the missing letter is the `V` of `FIVE`.

Elevator or stairs? A man lives on the 40th floor of a high-rise building that has elevators as well as stairs. Over time, his neighbors start noticing a bizarre habit of his, involving rides in the elevators. Entering the building from the street, he goes into the first available elevator. If there are others in the elevator, he stays there until reaching his floor, but if the elevator is otherwise empty, he always leaves the elevator at the 15th floor and climbs the stairs the rest of the way. When finally confronted about this strange behavior, he admits it, but gives a simple, reasonable explanation. What is it?

Solution. He is very short, only 4 feet 6 inches, so he can only reach the elevator buttons up to the 15th floor. When there are others in the elevators, they help him, but being by himself, he must quit the elevator on the 15th floor and climb the stairs all the way to the 40th floor.

Missing number. What is the missing number in the following

$$\begin{aligned} 5 + 1 + 6 &= 7 \\ 5 + 1 + 642 &= 70 \\ 4 + 6 &= ? \end{aligned}$$

Solution. $5 = s$, $+ = e$, $6 = n$, which should be enough data for you to guess that $4 = t$, implying that $4 + 6 = 10$.

There is no free lunch. This true story took place long, long ago, when lunch prices were much more reasonable. On one of their many dates, Alfredo took Violetta to lunch and proposed that they flip a coin to decide who should pay (one of his many faux pas). However, Violetta proposed that she would pay the entire tab if Alfredo's part came to exactly \$1.15, and that he pays everything in all other cases. To make this short story more interesting, Alfredo tried really hard to adjust his order to \$1.15, but failed, even though the menu was clear and accurate, and the tax was clearly specified. He reluctantly paid the substantial check, admitting that he was outwitted by Violetta's superior mathematical skills. It took him a while to propose another lunch date.



Figure 9.14: Image by MS Copilot, Sept. 2025.

Solution. In their state of Missitucky, sales tax was then four cents on anything between \$0.76 and \$1.10, and five cents on higher purchases, making it impossible to have a tab of \$1.15.

A question mark. On the second evening after Christmas, a snow ball containing a rock shattered a front window at Ebenezer Scrooge's house. Rushing to the broken window, Scrooge managed to see three kids whom he immediately identified as the rascal Pecksniff trio, Max, Mark, and Matthew. The following morning, a handwritten, torn and dirty message was found stuck to his front door:

```
? Pecksniff and you'll find who did it
```

Scrooge never understood this message, but you should have no problem. Just try.

Solution. Just replace the "`?`" with "`question Mark`."

Curious cube. The three views of the cube of Figure 9.15(a) look familiar and innocent, but this cube is non-standard. What is on the face opposite the six?

Solution. The confusing point is that two faces have two dots each. Another hint is to try to imagine the two faces with 2 and 3 dots side by side. The complete cube is shown in Figure 9.15(b).

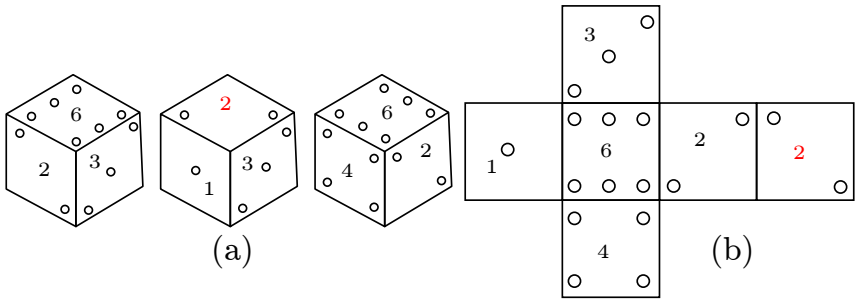


Figure 9.15: A Curious Cube.

Agatha, but not Christie. At 1:23 PM, Agatha, who has no idea of the time, checks her digital clock. It shows 1:24. "Oh shucks," she says, "the clock is again bad." How did she know that? Hint, the "PM" is significant.

Solution. The clock is not bad. It is simply set to 24-hour time.

Done by a mirror. While reading Lewis Carroll's *Through the Looking Glass*, Alice, the book's main character, started on her math homework. The first problem involved the product 50×1415.58 which, as we know, is 70779. However, when she looked at her calculator screen she saw to her horror the word `error`. What's the explanation?

Solution. She was still inside the book, where the 7-segment mirror image of 70779 is the word `error`.

Opposite of day tour. Steve had to wait for years to get from the telephone company the particular telephone number that he wanted. This 9-digit number has all distinct digits and starts with 061834. Consulting the left side of Table 9.16, which shows a standard telephone keypad, what are the remaining three digits?

1	2	3
4	5	6
7	8	9
$*$	0	#

1_3	2_8	3_5
4_6	5	6_2
7_9	8_4	9_7
$*$	0_1	#

Table 9.16: A Telephone Keypad.

Solution. The last three digits are 927. The right side of Table 9.16 shows that his nine digits are arranged on the keypad in a knight's tour.

Time traveler? Right after dinner, on the summer solstice, a time traveler in Iceland pushed a button on his time machine console. A display flashed, indicating

eighteen fifty-eight. A while later it indicated eighteen fifty-nine, followed by nineteen hundred. What did the display show next?

Solution. It showed 19:01. The display was simply a 24 hour digital clock.

Lucky Lindy. *The Spirit of St. Louis*, the airplane that took Charles Lindbergh from New York to Paris in 1927, had a single engine. The flight took more than 30 hours, and the engine worked flawlessly. Working non-stop for 30 hours was quite a feat for 1927 engines, and this puzzle deals with the question of whether it would have been safer for Lindbergh to have a two-engine plane.

We assume two identical engines, from the same maker, and in the same condition. An important assumption is that the airplane depends on both engines and is doomed if one of them fails.

Under such assumptions, would Lindbergh have been safer if he had two engines?



Figure 9.17: Image by MS Copilot, Sept. 2025.

Solution. Stripped to its essentials, the question is, if each engine has a 10% chance of quitting during 30 hours of continuous operation, what is the chance that one engine or the other would quit during this time period? Equation (7.2) deals with ORing probabilities, and applying it to our case produces the result $P(0.1 \text{ or } 0.1) = 1 - (1 - 0.1)(1 - 0.1) = 1 - 0.81 = 0.19$, or close to 20%. Lindbergh would be less safe in this situation.

This is why airplane manufacturers always make sure that a multi-engine plane can still fly even if several engines fail. If we assume that a two-engine plane can still fly on one engine, then the probability that both would fail is the small $0.1 \times 0.1 = 0.01$ or 1%, very safe.

`V` for victory? What letter, apart from `W`, can be substituted for the `?` in order to complete the natural sequence of letters in `V?XYZ` ?

Solution. The letter V, because the string "VV" may resemble a W in certain fonts. Thus WXYZ.

Friendships, odd and even. A number of people find themselves together. Necessarily, each of them knows some, perhaps zero, others and doesn't know the rest. We consider two people friends if each knows the other. Some people have an odd number of friends, while the rest have an even number (perhaps zero) of friends. Prove that the number of people who have an odd number of friends is itself even.

Solution. We start by asking each person i how many friends n_i they have. Denote the total number of friends by S. Since friendship is mutual, S must be even. The sum E of all the even n_i's is, of course, even. The sum of all the odd n_i's is denoted by D and we notice that $S = E + D$, which implies that D itself must be even. Thus, the number of people who have an odd number of friends is always even.

Rows of rose. Given 16 rose bushes, it is easy to plant them in a 4×4 array to end up with four rows, four columns, and two diagonals, a total of ten rows. Show how to plant those bushes so as to end up with fifteen rows.

Solution. Figure 9.18 shows how 16 red dots are distributed on 15 line segments which are arranged in three groups of five segments each.

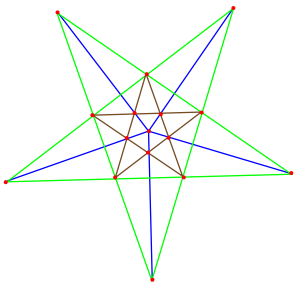


Figure 9.18: A Curious Cube.

A message. Agatha Christie has just received this message from miss Marple:

Meet me outside your Favorite pub tonight at 9

What time is this mysterious meeting?

Solution. Nine on the dot.

A perfect ten power. A fast, clever algorithm is needed to find a 10-digit pandigital integer that would also be a tenth power of another integer.

A pandigital number is an integer that in a given base has, among its significant digits, each digit used in the base at least once. See Page 163.

Solution. A clever approach is to first compute some numbers of the form a^{10} for small integers a, in order to find the 10-digit numbers that are tenth powers. For small integers a, such as 1, 2, and 3, it is easy to verify that a^{10} consists of fewer than 10 digits. Even $7^{10} = 282{,}475{,}249$ has only nine digits, but $8^{10} = 1{,}073{,}741{,}824$ and

$9^{10} = 3{,}486{,}784{,}401$ are 10-digits-long each, while $10^{10} = 10{,}000{,}000{,}000$ is 11 digits long. Hence, there is no such pandigital number.

The Two doors logic puzzle. You are in a cell with two doors, F and B. One leads to certain death and the other leads to freedom. A guard is stationed by each door, and you are allowed to choose a guard and ask one question. The point is that guard T always tells the truth, while guard L is a persistent liar, but you don't know which is which. What question would reveal the destinations behind the doors?

Solution. The simplest question would be: "which door is this?" but it is easy to see that this is useless. Suppose the safe door is F and we ask guard T. He would truthfully say "door F," but if we ask L, he would lie and say B. Since we don't know who is the lier, this question reveals nothing.

Now assume that we point at door F and ask "is this the safe door?" If F was the safe door, guard T would say "yes" and guard L would say "no." If B was the safe door, guard T would say "no" and guard L would say "yes." This question would provide no information either.

We now realize that any correct question must reveal two unknowns (1) which guard is lying and (2) which door is safe. The right question should somehow force the guards to respond identically and indirectly point us to safety.

A possible question is "what would the other guard say, if I asked him which door is safe?" Suppose the safe door is F and we asked Guard T. He knows that guard L would lie and answer "door B," so T would truthfully say "the other guard would say door B." Guard L, however, is a lier. He knows that T would truthfully say "door F," so if we ask L this question he would lie and say "the other guard would say door B." We get the same response from either guard, but this time we know that this response is wrong and door B is not the safe one. Clever!

If your head is still clear, you might want to analyze the following variation. There is only one guard, but he always lies (or he is always truthful). Your clever, two-part question is "If I were to ask you which door leads to safety, which door would you indicate?"

If the guard always lies, he would first consider what he would answer to "which door leads to safety?" If he was asked that question he would lie and indicate the dangerous door. However, to answer the question that he was actually asked, he has to lie, so would point to the safe door.

This, for me, is too much logic for one day, but if you want to continue, consider the variation where there is one guard, he lies half the time, but he would answer three questions.

How to win a lottery. Lottie, a young lottery enthusiast, asks: "Dad, what is this week's winning lottery number?"

Her dad's surprising answer: "Look at Table 9.19a with the six 5-digit recent winning numbers. Incredibly, and contrary to all the known rules of probability, the five digits of the current winner appear in the five columns of this table, and in their correct positions."

"Wow, how unbelievable," says Lottie, "but I still cannot figure out the number."

Saying this, Lottie turns around to show the table to her mother, who examines it and, surprise surprise, declares "dad is right, I also see the same five digits, each in its

0	9	9	0	6
0	I	6	9	I
8	9	6	8	9
0	9	6	0	9
I	9	6	8	9
8	6	8	8	9

(a)

6	8	8	9	8
6	8	9	6	I
6	0	9	6	0
6	8	9	6	8
I	6	9	I	0
9	0	6	6	0

(b)

Table 9.19: Six 5-digit Winning Numbers.

correct column. I cannot believe my eyes. Here is the number . . . "

"Don't," says Lottie, "I can see it now. It's so obvious."

Is it obvious to you too?

Solution. By turning around, Lottie showed her mother the table turned upside-down, as in Table 9.19b. Comparing the two tables showed Lottie that the winning number was 16891. Unfortunately, it is too late for her to benefit from this knowledge.

Multiplying by rotating. Here are two problems involving multiplication that is also a rotation. (1) Find the five digits a, b, c, d, and e that satisfy the relation $abcde7 \times 5 = 7abcde$. (2) Similarly, solve for $1abcde \times 3 = abcde1$, for the same five digits!

Solution. For part (1), we figure digit by digit from right (least-significant) to left (most-significant). The relation $5 \times 7 = 35$ tells us that $e = 5$ with 3 to carry. This also implies that $5e + 3 = 5 \times 5 + 3 = 28$, which makes it clear that $d = 8$, with 2 to carry. Continuing three more steps in this way, we easily discover that $c = 2$, $b = 4$, and $a = 1$. The solution is $142,857 \times 5 = 714,285$.

Part (2) is solved differently, but as easily. We denote by x the five-digit number $abcde$. This reduces the problem to the simple equation $(100,000 + x) \times 3 = 10x + 1$ whose solution is $x = 42,857$. Thus, $142,857 \times 3 = 428,571$. A lucky number indeed.

Hairy ones in Bucharest. Among a vast number of facts and useful information, Wikipedia tells us that the 2021 population of Bucharest, the beautiful capital and largest city of Romania, was 1,716,961. What it doesn't tell us is the following, unexpected question, implied by Figure 9.20 namely, are there in Bucharest two people who have the same number of hairs on their heads?

Solution. The obvious, trivial answer is Yes. There are many people in Bucharest, as well as in other places, who have the same number of hairs, namely zero. However, ignoring bald people, we present a simple, surprising answer.

Website `https://medicistanbul.net` claims that the average number of hairs on the heads of Caucasians is 150,000, implying that the number of individual hairs may range from 0 to 150,000. Employing the pigeon hole principle, we imagine a set of 150,001 boxes, where in box n we place all those inhabitants of Bucharest who have n hairs.

Since there are about 1.7 million residents, most boxes would be packed, illustrating the fact that for any hair count, there are many Bucharestians with the same number



Figure 9.20: Image by MS Copilot, Sept. 2025.

of hairs. The only way for the residents of this beautiful city to have unique numbers of hairs is to have one person in each box, but the difference between 150,001 and 1.7 million precludes this possibility.

Even more. If there were at most 11 people in any box (i.e., with the same number of hairs), the city would have at most $11 \times 150{,}001 = 1{,}650{,}011$ residents, but since it has about 1.7 million, there must be at least 12 people with n hairs for any $n \in [0, 150{,}000]$; a fact of great importance to Romanian barbers.

One plus one. We generally agree that one plus one equals two and that one plus four equals five, and these facts remain true if you substitute digits for the letters in the following examples. What are the digits?

$$\begin{array}{r} \text{ONE} \\ +\ \text{ONE} \\ \hline \text{TWO} \end{array} \qquad\qquad \begin{array}{r} \text{ONE} \\ +\ \text{FOUR} \\ \hline \text{FIVE} \end{array}$$

Solution. There are no special tricks here, just some thinking and several tries should produce the following:

$$\begin{array}{r} 286 \\ +\ 286 \\ \hline 572 \end{array} \qquad\qquad \begin{array}{r} 286 \\ +\ 3210 \\ \hline 3496 \end{array}$$

Toy soldiers. Little Soldi likes to play with lead soldiers. Before going to bed he bragged to his father that he managed at arrange his soldiers in two perfect rectangles. The first with 24 even rows and five soldiers left, and the second in 15 even rows with 16 soldiers put aside. The father, a mathematician, immediately realized that these two configurations result in a contradiction, but all he said to Soldi was "this is great, son.," and kissed him good night."

Before you, the reader, go to bed and say good night to me, spend a few minutes on this, won't you?

Solution. Assume that there are n soldiers. Their initial configuration was $n - 5$ soldiers in 24 even rows, which is why $n - 5$ must be divisible by 24. Later, $n - 16$

soldiers stood in 15 even rows, so $n-16$ should be divisible by 15. The factors of 15 are 3 and 5, and those of 24 are 2, 3, 4, 6, and 8. The only common factor is therefore 3. Notice that the difference of two numbers that are divisible by 3 is also divisible by 3, but the difference $(n-5)-(n-16)$ equals 11, which is not divisible by 3.

A colored cube. Figure 9.21 shows three views of the same cube, with its sides colored. What color is the bottom face of view 1?

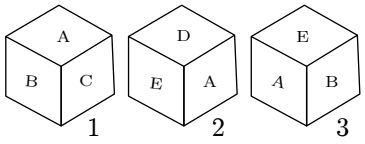


Figure 9.21: A Riddle Cube.

Solution. A thorough analysis of the three views should convince the reader that the bottom face of view 1 is A, same as the top face.

A clock problem. Grandpa Spock had a clock that he often forgot to wind. Yesterday evening, as he left the house on his way to his favorite pub, he noticed that his clock had stopped. He wound it, started it, and left. Coming back from the pub real late, he immediately set the clock accurately, without looking or checking anything. I still cannot figure out how he did that. Can you?

Solution. Ignoring the forgetful Spock, we strip the problem to its essentials. We denote by X the time the clock showed when Spock left the house; by A and D, the times he arrived at and departed the pub, respectively; and by Y, the time the clock showed when he returned home. We assume that he took time t to walk to the pub and the same time walking back (thankfully, no drunk).

Spock was therefore away for time period $(Y-X)$, of which he spent time $(D-K)$ at the pub. His total walking time of $2t$ must therefore equal the difference $(Y-X)-(D-K)$. This is enough data to compute t, and once t is known, his time of arrival home was $D+t$.

Milking a cow. How many legs are there when a cow is being milked?

Solution. Why, six.

Wrong. There are nine, because the person milking sits on a three-legged stool.

USE IT OR LOSE IT. In how many ways is it possible to read USE IT OR LOSE IT in table 9.22(a)?

Solution. Table 9.22(b) lists the number of ways to reach each location in part (a) of the table. For example, each of the two locations marked 3 in table (b) can be reached in three different ways when starting from the top. This is because the number of paths to reach a letter is the sum of the numbers of paths that its two parents can be reached. Thus, the total number of ways to meander in the table through the phrase `USEITORLOSEIT` is 924.

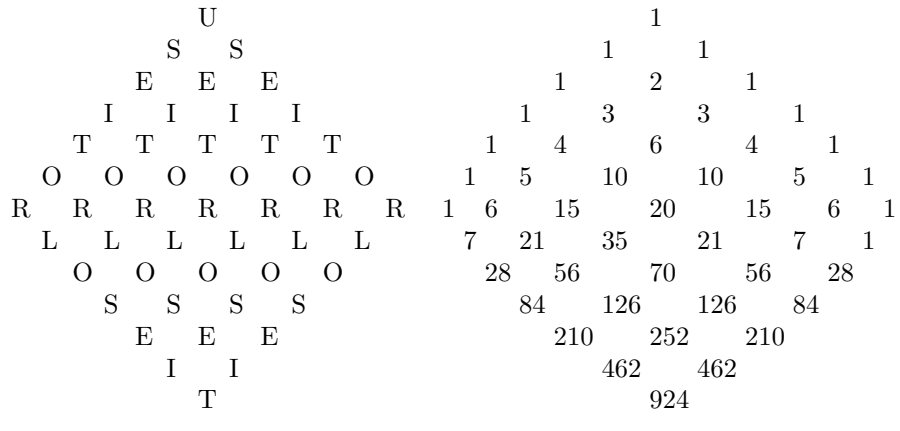


Table 9.22: USE IT OR LOSE IT.

Another Hilbert hotel? Seven tired hunters arrive late to a country inn, looking for rooms. "But we insist on having separate rooms," said their leader.

"Oh," said the innkeeper. "There are only six rooms here, but don't worry, I have my tricks of the trade. Follow me."

He invites the first guest to room 1, and asks another to wait a few minutes in the same room. He then brings the third guest into room 2, the fourth guest into room 3, the fifth guest into room 4, and the sixth guest into room 5. Then, he returns to the seventh guest in room 1 and invites him into room 6. The following morning he receives seven payments for his six rooms and gives thanks to Hilbert.

Solution. The guest being asked to wait a few minutes in room 1 was the second guest, not the seventh one.

First and Last Day of the Year. In 2025, both January 1st and December 31st fell on a Wednesday. In what other years is it true that the 1st of January and the 31st of December fall on the same day of the week?

Solution. This is true on non-leap years. Such a year has 365 days, a number that equals 52 weeks plus an extra day ($365 = 52 \times 7 + 1$). If January 1 of a year falls on a Wednesday, then each of the 52 weeks starts on a Wednesday and ends on a Tuesday, causing the extra day to be Wednesday. A leap year has 366 days, the extra day being February 29. Thus, if a leap year starts on a Wednesday, it will end on a Thursday.

Cats problem. A pet store has several dozen cats for sale. Some have blue eyes, while the others have green eyes. Some have gray coats, while the others have silver coats. A buyer wants two cats that together will have the two eye colors and two coat colors. Is it always possible to find two such cats?

Solution. Start by classifying the cats into four groups:

1. blue eyes and gray coat,
2. blue eyes and silver coat,

3. green eyes and gray coat,
4. green eyes and silver coat.

If group 1 is not empty, choose a cat from it. If group 4 is also non-empty, choose a cat from it and we are done. If group 4 is empty, then group 3 (green eyes) must be non-empty and group 2 (silver coats) must also be non-empty. Thus, the store can always satisfy this peculiar customer.

Jealous Husbands. Three couples traveling together arrive at a river. Only a single two-person boat is available for crossing, but the husbands are narrow-minded, jealous, and suspicious. None would leave his wife alone with another man even for a minute. If I were you, I would avoid those people, but you being you, may allow them to cross. So tell them how to do this.

Solution. We denote the three husbands by A, B, and C, and their wives by X, Y, and Z. Their needless suspicions require the following 11 steps:

1. A and X cross over;
2. A returns alone with the boat;
3. Y and Z cross over;
4. X crosses back alone;
5. B and C cross over;
6. B and Y cross back;
7. A and B cross over;
8. Z crosses back;
9. X and Y cross over;
10. Y crosses back;
11. Finally, Y and Z cross over if they are not too tired.

These are the wages of jealousy.

A winning strategy. The following two-player game has a simple winning strategy that allows the first player A to always win. A chooses an integer n between one and 10 and tells it to B, who then chooses an integer m in the same interval, and says the value of the sum $n + m$ to A. The two players alternate in this way, producing larger and larger sums, and the winner is the one who reaches 100 first.

What is the winning strategy?

Solution. When the current sum is in the interval $[90, 99]$, the next player can always reach 100 and is the obvious winner. Therefore, a player should try to reach the winning sum of 89. Once player A can say "sum is 89," he has effectively won, since B can only increase the sum to a value in the interval $[90, 99]$, which guarantees his loss.

Once this is clear, it becomes obvious that the way to reach 89 is to first reach a sum of 78, and the best way to do this is to first reach a sum of 67, and so on, through 56, 45, 34, 23, 12, and 1.

Therefore, the starting player A can always win if he chooses 1 as his first number. Even more. It is easy to see that 1 is the only winning initial number, because any other choice would enable B to bring his sum to the winning 12, from which A has no escape.

The magic of 1089. Pick a three-digit number $x = abc$ such that a and c differ by 2 or more. Reverse the digits to obtain the mirror image $y = cba$. Subtract $z = |x-y|$ to obtain a 3-digit integer z, and add z to its mirror image. Show why the result is 1089.

Solution. It is clear that $x = 100a + 10b + c$ and $y = 100c + 10b + a$, so their difference is $z = 100(a - c) + (c - a)$. We assume, without loss of generality, that $a > c$, but this means that $c - a$ is negative, which is why we rewrite z in the form $100(a-c)+(c-a) = 100(a-c-1)+10\cdot 9+(10+c-a)$. Adding z to its mirror image produces

$$\begin{aligned}&100(a-c-1)+10\cdot 9+(10+c-a)+100(10+c-a)+10\cdot 9+(a-c-1)\\&\quad=-100+90+10+1000+90-1=1089.\end{aligned}$$

Repeating this process for $a = c$ results in 0, and doing it for $|a - c| = 1$ produces 99.

The secret of 32. The number 32 has many factors, but did you know that it can also be represented as an expression that uses only the number 1 and the four arithmetic operations?

Your task is find nine expressions of 32, each employing just one of the decimal digits 1 through 9.

Solution. Here are nine solutions

$$\begin{aligned}32&=11\times(1+1+1)-1\\&=2\times 2^{(2+2)}\\&=33-3/3\\&=4\times(4+4)\\&=((5+5)/5)^5\\&=6\times 6-6+(6+6)/6\\&=7+7+7+77/7\\&=8+8+8+8\\&=9+(99+99+9)/9.\end{aligned}$$

See also Page 72.

200 trillion. Find the only integer between 2 and $2 \cdot 10^{14}$ that is a (nontrivial) square, cube, and fifth power. That's your entire task.

Solution. A number that is both a square and a cube must also be a sixth power as is, for example, 64. It equals both 8^2 and 4^3, which makes it also 2^6. Since $2 \times 3 \times 5 = 30$, our mystery number must be of the form x^{30} and must be $\leq 2 \cdot 10^{14}$. We start checking integers from 2 (because 1 is a trivial choice). The first candidate is $2^{30} = 1{,}073{,}741{,}824$. It is close to the limit and it is the winner, because the second choice, $3^{30} = 2.0589113209464900000 \cdot 10^{14}$ is too big. No need to check trillions of numbers.

Seven plus eight. Operating on numbers in base-13, it is true that $7 + 8 = 12$, but what about the following sum

$$\begin{array}{r}\text{SEVEN}\\ \text{EIGHT}\\ \hline \text{TWELVE}\end{array}$$

Solution. The following solutions may not be the only ones

$$\begin{array}{r} 69298 \\ 90431 \\ \hline 159729 \end{array} \quad \text{or} \quad \begin{array}{r} 85254 \\ 50671 \\ \hline 135925 \end{array} \quad \text{or} \quad \begin{array}{r} 63732 \\ 39841 \\ \hline 103573. \end{array}$$

A special knight's tour. An important piece in the game of chess is the knight, whose valid moves are unintuitive. It can move two squares in one direction (horizontally or vertically), followed by one square perpendicular to that direction.

The knight's tour problem is finding a sequence of moves for a knight on an $n \times n$ board, such that it visits every square exactly once. There is a closed version of the tour, where the knight returns to its starting position.

Consider the special version of the tour where $n = 8$, the knight starts at the lower-left corner, and ends at the upper-right corner. Is such a tour possible?



Figure 9.23: Image by MS Copilot, Sept. 2025.

Solution. The answer is no, and here is the proof. The board consists of $8 \times 8 = 64$ squares, painted alternately black and white. The knight starts at the lower-left corner which is white, and it has to perform 63 steps. It is easy to verify that at each move it ends up at a square of the opposite color. Thus, after 63 steps it must end up on a black square, but the upper-right square is white.

The well-known 15 puzzle is analyzed similarly. See Wikipedia article `15_puzzle`.

General knight's tour. Try your hand at an 8×8 knight's tour, where the start and end points can be anywhere on the board.

Solution. Table 9.24 shows one of many solutions. Wikipedia article `Knight's_tour` discusses approaches to solving this kind of problem. See also the similar Table 10.8. These tours are also magic squares where the sum of the numbers in each row and in each column is 260.

Rook Walk. A rook stands on the lower-left corner of a chess board. Your task is to move it to the upper-right corner so that it crosses each square exactly once, similar to a knight's tour.

50	11	24	63	14	37	26	35
23	62	51	12	25	34	15	38
10	49	64	21	40	13	36	27
61	22	9	52	33	28	39	16
48	7	60	1	20	41	54	29
59	4	45	8	53	32	17	42
6	47	2	57	44	19	30	55
3	58	5	46	31	56	43	18

Table 9.24: An 8×8 Knight's Tour Magic Square.

Solution. In each step, a rook can be moved horizontally, vertically, or diagonally, but only one square. In order to cover the entire 64-square board, the rook has to perform 63 steps, but any odd number of steps moves it from a square to an opposite-color square. Thus, the problem is unsolvable.

Macalester Problem #1321. Since 1993, Stan Wagon has been in charge of the Macalester College Problem of the Week. His problem #1321 has to do with powers of 2. Given a set of n unique integers, it is possible to choose two out of them in $\binom{n}{2}$ ways. The Mathematica statement `Table[Binomial[n,2],{n,10}]` prints the first 10 values of $\binom{n}{2}$ `{0,1,3,6,10,15,21,28,36,45}` for positive integers n.

Given a value for n, the problem is to choose a set of n unique, unordered, signed integers such that the sum of every pair of them will be a positive power of 2. For $n = 2$ this is trivial. The pairs (0,2), (1,3), and (2,4), are only some of the many choices. For $n = 3$, the triplet $(-1, 3, 5)$ produces the three sums $-1 + 3 = 2$, $-2 + 5 = 3$, and $3 + 5 = 8$, which satisfy the requirement. Your task is to try your hand with $n = 4$.

Solution. No perfect solution has ever been found. Nor is it currently known whether there is such a solution. The best that is known today is a set such as $(-3, -1, 3, 5)$ where four out of the possible six pairs sum to a positive power of 2. This is one of many examples of problems in number theory that are easy to state and understand, but resist repeated attempts at a solution or a counter example.

Grandfather and Grandson. Mathematicians are often considered boring, but our interest here is about one whose age in 1932 equaled the number formed by the two last digits of his birth year. What makes his story worth including here is that the same was true about his grandfather in 1932. When where they born?

Solution. We assume that the mathematician was born in the 20th century, while his grandfather (also a mathematician?) was born in the 19th century. Denote by x the two-digit number formed by the last two digits of the mathematician's birth year, then he was born in year $1900 + x$. In 1932 his age was x years, so we can write $(1900 + x) + x = 1932$ which is solved to $x = 16$ and a birth year of 1916.

The analysis in the grandfather's case is similar. He was born in year $1800 + y$, so $(1800 + y) + y = 1932$, which translates to $y = 66$ and a birth year of 1866.

A square mathematician. Augustus De Morgan, a well-known 19th century English mathematician, told his birth year in the following puzzle: "I was born in this

century, and was in the fortunate position that in the year x^2 I was just x years old." When was he born?

Solution. The 19th century included the years 1801 through 1900, an interval that includes only one square number $1849 = 43^2$. De Morgan therefore turned 43 in 1849, implying that he was born in 1806, coinciding with the return of the Lewis and Clark Expedition, as well as the end of the Holy Roman Empire.

The Mountaineer. Ben Harari, an experienced mountaineer starts climbing mount Indefatigable at 8 AM and reaches the summit at 6 PM. He sets his tent, eats, rests, and sleeps. The following morning, again at 8 AM, he starts down the mountain, retracing his path, and reaches his start point at 4 PM. Show that there must be a point somewhere along his path which he reaches at the same time on the two days.

Solution. This is based on simple logic instead of calculus. Assume that there are two climbers and both start at the same time, taking the same route, one from below and the other from the top. The point where they meet is our solution. Simple!

A Poem about Puzzles

In numbers pure, where logic reigns,
Mathematical puzzles spark our brains,
With elegance, they twist and weave,
Revealing truths that we perceive.

The primes, they whisper secrets old,
In patterns hidden, stories told,
A dance of digits, sharp and clear,
In every theorem, they appear.

Fibonacci's spiral flows,
In nature's way, the sequence grows,
From flowers blooming, stars aligned,
In puzzles vast, their paths we find.

Magic squares with sums divine,
In every row, the numbers shine,
A balance struck in every grid,
Where logic's light can never be hid.

Pythagorean paths unfold,
In triangles, the truth is told,
A squared plus B squared, simple, true,
In every right, the proof comes through.

Sudoku's grids, a numbered field,
Where strategy and skill are wield,
Each row and column, box defined,
A symphony of the ordered mind.

The mysteries of pi, unbound,
In endless digits, truths are found,
A circle's story, never done,
In puzzles, endless as the sun.

Mathematical puzzles, bright,
Illuminate the mind's great light,
In every problem, beauty's seen,
A quest for knowledge, pure and keen.

A jigsaw spreads its colored wings,
Each fragment part of wondrous things,
A landscape bright, a sea so vast,
A journey through the pieces cast.

For in these challenges, we find,
A deeper joy, a sharper mind,
In numbers vast, our thoughts set free,
To solve, to wonder, endlessly.

By `chatGPT 4o`, June 2024.

I do the 'New York Times' crossword puzzle every morning to keep the old grey matter ticking.

—Carol Burnett.



10
Miscellaneous topics

10.1 The Gamma Function

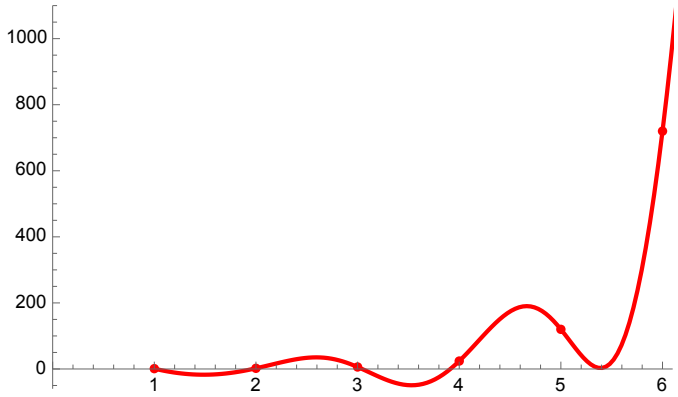
Factorials grow very quickly, but mathematicians, being both curious and devious, have found ways to define super factorial functions that often produce very large outputs for very small inputs. See reference [SuperFactorial 21] for funny examples.

◇ **Exercise 10.1:** Try to come up with a practical example of a very large factorial.

The gamma function. The factorial function is defined for positive integers, which is why it is natural to try to extend it to any real or complex numbers. When considering the question what is the factorial of 1.5, an immediate idea that may come to mind is to interpolate and locate the point midway between 1! and 2!. This can be done by computing the smooth curve that passes through the points $(k, k!)$ for k values from 1 to a certain n, as shown in Figure 10.1. However, the figure makes it clear that such naive interpolation, which is done with the Lagrange polynomial, while smooth, oscillates widely and does not provide the correct values.

An alternative is to compute a complete interpolation of n factorial points for any n by means of individual segments, each a simple polynomial interpolating between two points $(k, k!)$ and $(k + 1, (k + 1)!)$, and connecting the segments smoothly, such that the entire interpolation looks like a single smooth curve. Several such spline-based interpolation methods are described in detail in chapters 8–14 of [Salomon 11].

That said, it is clear that the correct approach to extending the factorial is to preserve its important "spirit" which is its recursive definition. The value $n!$ is the product of its predecessor $(n - 1)!$ and n. Thus, when trying to extend the definition of factorial to nonintegers, to negative numbers, and even to complex numbers, it is essential to preserve this property, so as to have $z!$ equal $z(z - 1)!$. This problem was




```
ListLinePlot[Evaluate[Table[k!, {k, 1, 7}]], InterpolationOrder -> 6,
 PlotMarkers -> {Automatic}, PlotStyle -> Directive[Red],
 PlotRange -> {{0, 6.1}, {-60, 1100}}]
```

Figure 10.1: Smooth Interpolation of Factorials.

initially solved in the 1720's by Daniel Bernoulli who came up with the horribly-looking and hard-to-compute expression

$$x! = \lim_{n\to\infty} \left(n + 1 + \frac{x}{2}\right)^{x-1} \Pi_{k=1}^{n} \frac{k+1}{k+x}.$$

In 1729, this impractical expression was superseded by our friend Leonhard Euler, who proposed the much simpler integral

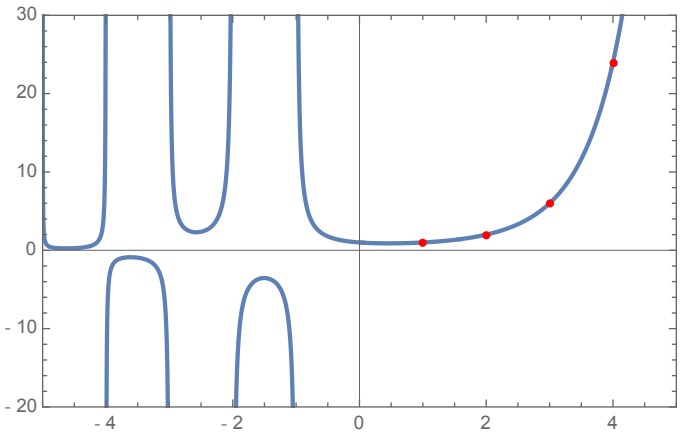
$$n! = \int_0^1 (-\ln s)^n \, ds,$$

which was later named the gamma (Γ) function and was modified by the change of variable $t = -\ln s$ to

$$\Gamma(z) = \int_0^\infty \frac{t^{z-1}}{e^t}\, dt \quad \text{or equivalently } \Gamma(z+1) = \int_0^\infty \frac{t^z}{e^t}\, dt = \int_0^\infty x^z e^{-x}\, dx. \qquad (10.1)$$

This function passes through the points $(n, n!)$ that correspond to familiar factorials of the positive integers. It is never zero, and is undefined for the negative integers, but not in between them. There are other ways to extend the basic factorial (reference [gamma.history 24]), but the Gamma function has become the standard.

As Figure 10.2 makes clear, if we adopt Gamma as the extension of factorial, then $z!$ for any complex number z is simply $\Gamma(z+1)$. For example $\pi! = \Gamma(\pi+1) = 7.18808$, $(1/2)! = \Gamma(1+1/2) = \sqrt{\pi}/2 = 0.886227$, and $(-1/2)! = \Gamma(1-1/2) = \sqrt{\pi} = 1.77245$.




```
L1 = Plot[Gamma[x + 1], {x, -5, 5}, Frame -> True,
   PlotRange -> {{-5, 5}, {-20, 30}}];
L2 = ListPlot[Gamma[Range[4] + 1], PlotRange -> {{-5, 5}, {-20, 30}},
   Axes -> False, PlotStyle -> Directive[Red]];
Show[{L1, L2}]
```

Figure 10.2: Euler Gamma Γ Function.

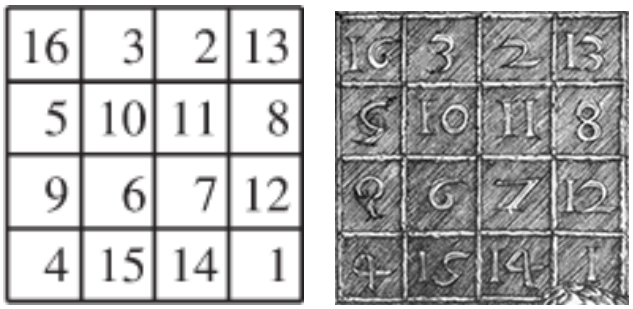
10.2 Magic Squares

A magic square is an $n \times n$ array of integers arranged such that the sum of the n numbers in any horizontal, vertical, and the two main diagonals is the same. The obvious choice is simply to have the same number in every position, but this is trivial and leads to the conclusion that a *real* magic square must consist of the distinct integers 1, 2, ..., n^2, and a magic square that has repetitions of the same number should be considered somewhat less than perfect, perhaps semi-magic, pseudo-magic, or partly-magic.

One of the best known magic squares is included in Melencolia I, a large engraving by the German Renaissance artist Albrecht Dürer. Its date, 1514, is part of the square and can be seen in the two middle cells of the bottom row of Figure 10.3. The square sums to 34, i.e., each of its rows, columns, diagonals, corners, and quadrants adds to 34. Even more. Any of its entries added to its symmetric opposite yields 17 (e.g., 15+2, 9+8). Magical indeed!

Perhaps the most important structure (and certainly the most important tourist attraction) in Barcelona is the Sagrada Familia (sacred family) cathedral, Antoni Gaudi's monumental project. It was started in 1883 and should be completed in 2026. Our interest in this unique structure is an inconspicuous 4×4 magic square (Figure 10.4), located on the Passion facade. Each of its rows, columns, diagonals, and 2×2 subsquares adds to 33. This isn't a perfect magic square because it doesn't include all the integers from 1 to 16 and because each of 14 and 10 appears twice.

Figure 10.5 illustrates a 4×4 Magic square that sums to 264. Every row and column, the two diagonals, and every 2×2 grid in this square adds to 264. Amazingly, it can also be turned upside down, and even reflected to its mirror image, while retaining its



16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

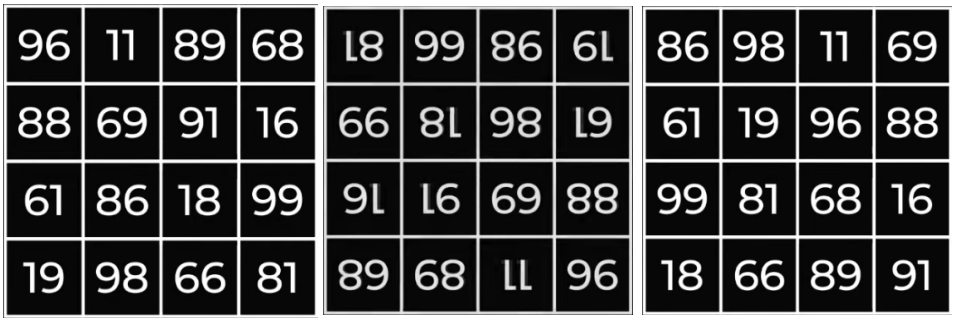
Figure 10.3: The Dürer Melancholia Magic square.



1	14	14	4
11	7	6	9
8	10	10	5
13	2	3	15

Figure 10.4: Sagrada Familia Magic Square.

magic. This feature becomes more noticeable when a 7-segment font is used to print the numbers.



96	11	89	68
88	69	91	16
61	86	18	99
19	98	66	81

18	99	86	61
66	81	98	19
91	16	69	88
89	68	11	96

86	98	11	69
61	19	96	88
99	81	68	16
18	66	89	91

Figure 10.5: A Magical Square with Sum 264.

The following is not strictly a magic square because some of its entries may not be unique, but it serves as a nice mathematical game. For any integer n, the following 4×4

square appears to be magic, with a sum of n.

$$\begin{array}{cccc} n-20 & 1 & 12 & 7 \\ 11 & 8 & n-21 & 2 \\ 5 & 10 & 3 & n-18 \\ 4 & n-19 & 6 & 9 \end{array}$$

At a party, it is best to use it with small sums. The magician may say, for example, state your age, and then quickly construct a square that sums to that age. Otherwise, someone may wonder how a magic square with a sum of, say, 1000 contains mostly small numbers. A careful examination immediately reveals the secret of this square. Each row, column, diagonal, and 2×2 region sums up to n. Clever!

Mathematics is rigorous, but mathematicians are human and many like to stretch the strict limits of their profession in order to have a good time. The following square does not follow the original definition but in some sense it is doubly magical, because each entry is a triple (Figure 10.6). The bottom component b of a triplet is part of a magic square, the middle part m is an alphabetic string of length b, and the top components t is the integer described by m, as well as part of another magic square. The triple $(t, m, b) = (12, \text{twelve}, 6)$ is an example. Several such squares are demonstrated in reference [magic.square.tri 23] by Matt Parker, of youtube channel `Stand-up Maths`.



Figure 10.6: A Double Magic Square.

Reference [magic.square3 23] shows why there is only one 3×3 magic square, and how to construct it. Similarly, reference [magic.square9 23] describes the Lo-Shu method of constructing a 9×9 magic square.

Table 10.7 is a 3×3 magic square which is also bimagic. The squares of its elements form their own magic square.

9	2	7
4	6	8
5	10	3

9^2	2^2	7^2
4^2	6^2	8^2
5^2	10^2	3^2

Table 10.7: A Bimagic Square.



One of the pieces of the game of chess is a knight (also known as a horse in some languages). It moves two squares vertically and one square horizontally, or two squares horizontally and one square vertically, jumping over other pieces as needed. Mathematicians, being curious, have pondered the following problem since the 9th century A.D. Is it possible for a chess knight to move 64 times, covering the entire chess board and visiting each square exactly once? This has become known as the knight tour problem.

In 1991 it was proved that for any $m \times n$ board with $m \leq n$, a closed knight's tour is always possible unless one or more of these three conditions are met:

- m and n are both odd
- $m = 1$, 2, or 4
- $m = 3$ and $n = 4$, 6, or 8.

A knight's tour is considered closed if it ends on a square that is one knight's move from the beginning square (so that the knight could tour the board again immediately, following the same path). Otherwise, the tour is termed open.

Table 10.8(left) is an 8×8 knight's tour that is also a magic square where each row and column sum to 260. The sums of the diagonals, however, are different, which turns this into an example of a semimagic square. (See also Table 9.24.) The right side of the table is an interesting 10×10 knight's tour where only 99 squares are covered. This tour, which is not a magic square, was discovered by Georges Perec and used by him to arrange the 99 chapters of his novel *Life: A User's Manual* (1978).

1	48	31	50	33	16	63	18
30	51	46	3	62	19	14	35
47	2	49	32	15	34	17	64
52	29	4	45	20	61	36	13
5	44	25	56	9	40	21	60
28	53	8	41	24	57	12	37
43	6	55	26	39	10	59	22
54	27	42	7	58	23	38	11

59	83	15	10	57	48	7	52	45	54
97	11	58	82	16	9	46	55	6	51
84	60	96	14	47	56	49	8	53	44
12	98	81	86	95	17	28	43	50	5
61	85	13	18	27	79	94	4	41	30
99	70	26	80	87	1	42	29	93	3
25	62	88	69	19	36	78	2	31	40
71	65	20	23	89	68	34	37	77	92
63	24	66	73	35	22	90	75	39	32
	72	64	21	67	74	38	33	91	76

Table 10.8. Left: An 8×8 knight Tour Magic Square. Right: A $(100 - 1)$ Knight Tour.

Table 10.9 is the smallest trimagic square that can exist. Its sums are 870, but when its elements are squared, their rows, columns, and diagonals sum to 83,810, and

1	22	33	41	62	66	79	83	104	112	123	144
9	119	45	115	107	93	52	38	30	100	26	136
75	141	35	48	57	14	131	88	97	110	4	70
74	8	106	49	12	43	102	133	96	39	137	71
140	101	124	42	60	37	108	85	103	21	44	5
122	76	142	86	67	126	19	78	59	3	69	23
55	27	95	135	130	89	56	15	10	50	118	90
132	117	68	91	11	99	46	134	54	77	28	13
73	64	2	121	109	32	113	36	24	143	81	72
58	98	84	116	138	16	129	7	29	61	47	87
80	34	105	6	92	127	18	53	139	40	111	65
51	63	31	20	25	128	17	120	125	114	82	94

Table 10.9: A Trimagic Square.

when the same elements are cubed, the square remains magic, with sums of 9,082,800. A magical feat indeed.

Next, multiplicative magic squares. Yes, they do exist (Table 10.10 left, center), and there are even additive-multiplicative such squares (Table 10.10 right, where the sums are 465 and the products are 150,885,504,000).

9	2	7
4	6	8
5	10	3

2	9	12
36	6	1
3	4	18

126	66	50	90	48	1	84
20	70	16	54	189	110	6
100	2	22	98	36	72	135
96	60	81	4	10	49	165
3	63	30	176	120	45	28
99	180	14	25	7	108	32
21	24	252	18	55	80	15

Table 10.10: Multiplicative Magic Squares.

A somewhat-magic square? It is not hard to verify that the rows and columns depicted in the 4×4 square of Table 10.11 are not just regular, but have the following property. If you select one number from each row, making sure that no two numbers are from the same column, their sum would always be 34. The simplest example is the main diagonal, which produces $1+6+11+16 = 34$. This is certainly not a magic square, but we can perhaps consider it somewhat-magical.

A brute force way to explain this property is to check every possible combination of four numbers chosen from distinct rows and columns. There are four ways to choose a number from the top row, but once we have chosen it, there are only three ways to choose a number from the second row, and similarly, only two ways to choose one from row 3, and one way to choose from the bottom row. The total number of ways to choose four such numbers is therefore $4 \cdot 3 \cdot 2 \cdot 1 = 24$, not too big, but not too clever either. A cleverer approach is needed, one that would make it possible to predict the behavior of similar but larger squares.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Table 10.11: A Somewhat Magic Square.

We start by choosing the four numbers in the leftmost column. They add up to $1+5+9+13=28$, but they clearly constitute a wrong choice. We try to improve out choice by moving numbers to different columns. We leave the top-left number, 1, alone, and start moving the three remaining numbers so as to satisfy the somewhat-magical requirement. By moving any of these numbers to the second column, we would be adding 1 to the total of 28. Similarly, but moving any to the third column, we would be adding 2 to the total, and by moving any to the rightmost column, the total would grow by 3. Hence, moving the three numbers to three distinct columns would add $1+2+3=6$ to the total of 28, bringing it up to 34.

Next, we observe that moving to the right, column by column, in the square of Table 10.11, increases the value of the current number by 1, while moving down from row to row, increases the value of a number by 4. If we number the columns 1 through 4 from left to right and number the rows 1 through 4 top to bottom, we can summarize this by writing $1\times(\text{column})$ and $4\times(\text{row}-1)$.

Since the top-left number is 1, we conclude that the value of the number on row r column c is $4(r-1)+c=4r+c-4$. Our task is to choose four numbers in cells whose (unknown) coordinates are (r_i,c_i) for $i=1,2,3,4$. The sum of the numbers in those cells is

$$\sum_{i=1}^{4} 4r_i+c_i-4=4(r_1+r_2+r_3+r_4)+(c_1+c_2+c_3+c_4)-4\cdot 4,$$

But we know that the four rows must be different, which is why their sum must be $1+2+3+4=10$, and similarly for the columns. Thus, the sum of the four numbers, if we choose distinct rows and columns, is always $4\cdot 10+10-4\cdot 4=34$.

It is easy to see why rotating and transposing this square leaves the sum of 34 unchanged. It is not also obvious that a similarly-constructed $n\times n$ square would result in a sum of $\sum_{i=1}^{n} nr_i+c_i-n$.

Anti-magic squares may be the favorite of some square enthusiasts. Figure 10.12(left) shows such a square, where the sums of the rows, columns, and one diagonal are not equal, but range from 29 through 38. Each of the three sides of the magic triangle of the same figure has three numbers that add up to 9. Each of the five rows and 10 diagonals of the order-3 hexagon sums to 38. Each of the four circles and four diagonals sums to 132. Lots of magic is hidden even in non-square shapes.

Finally, a really magical magic square, by no other than the magical Ramanujan. Figure 10.13 shows that in addition to the rows, columns, diagonals, and four corners, this square contains many groups of four numbers that sum to 139. Magical indeed.

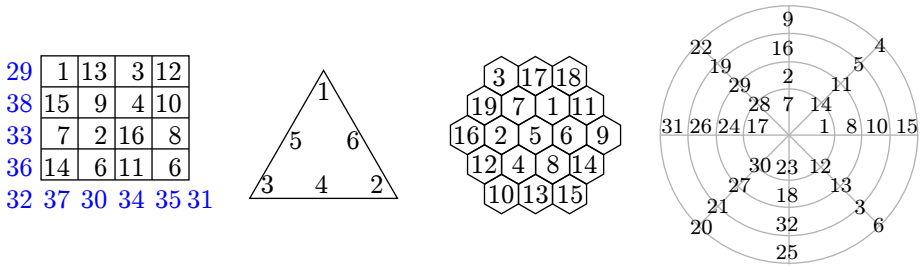


Figure 10.12: Magic Non-Squares.

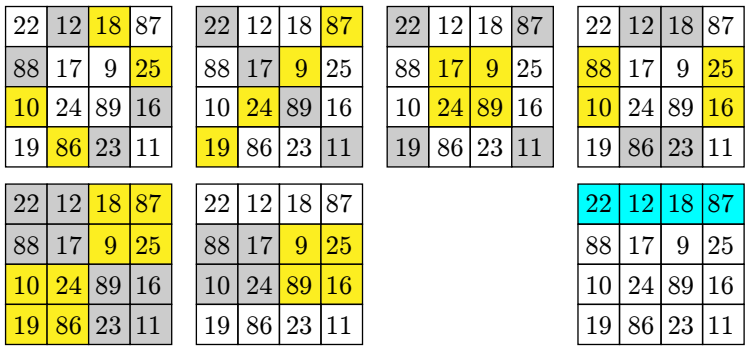


Figure 10.13: Ramanujan's Magic Square.

Even the top row, magically contains Ramanujan's birthdate, the 22nd of December, 1887; unbelievable!

10.3 Parking as a greedy problem

Finding a parking spot is a major problem in today's society. While traveling across Europe and crossing borders frequently, you may find yourself hearing foreign conversations and trying to decipher street signs written in unfamiliar languages. In these instances, the first words you may familiarize yourself with in a new language are often "No Parking!"

Large shopping centers are common in many countries and such centers often have vast parking lots organized in parallel lanes. It may be easy to find parking even in the busiest seasons, but human nature is such that we always try to find the parking spot nearest the front door, thereby turning parking into a greedy problem. It therefore happens that we get into a parking lane and drive slowly, looking left and right, trying to find an empty space. In such a lane it may be impossible to see much ahead, so when we see an available space we must quickly decide whether to park there or try for a nearer,



Figure 10.14: Image by chatGPT, Sept. 2025.

better spot. This is a decision that drivers in big cities face often, which explains why mathematicians have also shown an interest in it and have tried to analyze and solve it.

In order to treat this parking problem analytically, it has to be simplified and stated precisely. We therefore assume that a driver is trying to park close to a busy destination while driving along a one-way street that has many adjacent parking spaces on one side. The driver can read the number of a space as well as its state (empty or full) only when he gets to it. There is no backing up. The spaces are numbered, and the ideal one, the one nearest the destination, is #100. The driver starts looking for a space when he gets to around space 80; lower-numbered spaces are considered too far. He may overshoot space #100 and park at a higher-numbered space as a last resort.

Under such conditions, what could be a reasonable parking strategy? How can the driver estimate the general availability of spaces around his destination? Clearly, this problem is probabilistic, it depends on the density of parking at the time of arrival. If only 10% of the spaces are occupied, a driver would notice it right away and would be able to find a good spot without resorting to mathematics. We should therefore assume a high parking density, perhaps as high as 90%, and keep this in mind.

When you find an empty space, you should weigh the factors of parking there. Those factors are the (known) distance between the space and the ideal space (#100), and the (unknown) distance between this space and the next available space. Space #100 is a good spot to start. If this space is empty, there is no need for a strategy; you simple park there and that's it. If it is occupied, you must drive ahead and wait for the next unoccupied space, where you should better park, since you are already beyond #100.

Thus, if you are at #100 or beyond, your strategy is clear. Even more, in this case it is easy to estimate the expected distance between #100 and the next empty space. Assuming a 90% parking density, there is a 10% probability that space #100 would be available, implying a zero distance. In the 90% case where #100 is taken, the average number of spaces we would need to cover before finding an empty one is $10\% \cdot 0 + 90\% \cdot 10 = 9$ (the mean of a geometric random variable).

This takes care of the case where we are already at space #100. What if we are at

#99? If it is open, we better park there, ending up with a distance of 1. If we don't part at #99, we should move to #100, where the average distance is already known to be 9. Thus, there is a probability of 10% (combined with distance 1) that space #99 will be open, and a probability of 90% (combined with distance 9) that space #99 will be occupied. The average distance is therefore $10\% \cdot 1 + 90\% \cdot 9 = 8.2$.

Similarly, the situation at space #98 is a probability of 10% (combined with distance 2) that it will be open, and a probability of 90% (combined with distance 8.2) that it will be taken. The average distance in this case is $10\% \cdot 2 + 90\% \cdot 8.2 = 7.58$. Continuing in this way, the expected distances for spaces #97, #96, and #95 diminish steadily until, at space #94, the distance is 6.57. If this space is empty, we should park there, because the expected distance to the next empty space is 6.57, but the distance to our destination is 6.

The next spot to examine is #93. If we park there, the distance to our destination is 7, but if we continue to #94, the expected distance is the shorter 6.57. Thus, even if #93 is free, we should continue to #94 and examine it. We now realize that with a parking density of 90%, space #94 is our threshold. We should ignore any available paces before #94, but park at the first empty space following that.

This approach to estimating the expected parking distance is referred to as backward induction, because it goes backward from the ideal case, spot #100, to find the best strategies for lower-numbered spaces. This method is also an application of the more general approach of dynamic programming, a well-known, powerful optimization technique. The description here is based on reference [Optwhiz 23].

10.4 Error-Control Codes

The computer revolution that took place during the second half of the 20th century was mostly based on fast, inexpensive computers and on computer networks, specifically, the Internet. At the time of writing (late 2023), life in large parts of the world is inconceivable without computers and fast data communication. Computer users have gotten used to this way of life and are normally unaware of all the details that went into implementing the hardware and software that make it possible. The topic of this section is one of these details, namely the clever, sophisticated codes that detect and correct many of the errors in our data.

The sentence "the world is a noisy place," is often heard. It usually refers to the acoustic noise that we hear around us all the time, in crowded places, in restaurants, and on noisy streets, but there is another type of noise, noise that we don't hear but that pervades the entire world, even the quietest places. This noise causes machines to make mistakes, it garbles data transmissions, and in general degrades the performance of the tools and devices that we routinely use. Much of this noise is created by electromagnetic interference resulting from lightning, solar storms, solar flares, high-power radio, and malfunctioning or poorly designed electronic devices.

As a result, data sent inside a computer or between computers collects errors on its way. If the data is analog, such errors distort the data and garble its meaning to some degree. However, most of our current data is digital, where the corruption of even one bit may have serious effects. Medical data, both text and images, legal documents, and

commercial contracts are especially vulnerable to data corruption and must be protected. Considering the vast amount of data in the form of text and images that we all send, receive, and share, even a small percentage of errors would translate to many millions of bad documents received every day all over the world, which is why it is imperative to develop and use strong codes that can correct, or at least detect, many common errors.

It is the nature of digital data transmission that errors corrupt the elements of a transmitted data, but they do not delete or add any such elements. Because of the use of binary numbers in computers, the basic elements of our data are bits, and a bit can have only one of two values which are usually referred to as 0 and 1. This means that if we can identify a corrupt bit, we can easily correct it. Thus, the problem of error correction is essentially the task of identifying bad bits.

The mathematical branch of error-control codes (codes that detect or correct errors) is huge and includes many methods and algorithms, but it is based on a single, most-important idea, namely redundancy. Any error-control code works by somehow adding redundancy to the data, thereby making it more reliable, but also longer.

(Definition: In engineering, redundancy is the inclusion of extra, unnecessary components which are included only to automatically replace other components in case of failure.)

The idea of adding redundancy to digital data is the brainchild of Claude Shannon, who made it part of his 1948 revolutionary information theory. However, the similar idea of adding redundancy to natural languages has already occurred to people in the distant past, when those languages developed. Here is how this powerful technique works. European languages tend to have a small number of letters, normally around 20–30, from which a vast number of words can be constructed. The number of three-letter words that can be derived from the 26 letters of the English alphabet is $26^3 = 17576$ and the number of seven-letter words is $26^7 \approx 8.031$ billion. On the other hand, we use only about 50,000 words on a daily basis, many dictionaries include only about 100,000-250,000 words, and even the currently biggest dictionary, the authoritative web-based Wiktionary lists "only" 711,378 headwords (in 2023).

> A headword, lemma, or catchword is the word under which a set of related dictionary or encyclopedia entries appears. The headword is used to locate the entry, and dictates its alphabetical position. Depending on the size and nature of the dictionary or encyclopedia, the entry may include alternative meanings of the word, its etymology, pronunciation and inflections, compound words or phrases that contain the headword, and encyclopedic information about the concepts represented by the word.
>
> —From *Wikipedia*, 2023.

The huge parity between the vast number of words that can be created by a few letters and the actual number of words in any given language is the source of redundancy in the language.

This redundancy is one reason why we can often recognize a word when we hear only part of it. When we hear "?red" we can associate it with similar words in our language such as bread, dread, and Fred, and then choose the word that makes the most sense depending on the context. We may choose bread if we are in line in the bakery, Fred if we are talking to his brother, and dread if the conversation around us is about war or locks.

Thus, redundancy in a natural language comes to us naturally, but redundancy in digital data was a revolutionary concept, because we are conditioned against it. Most of the time, we try to reduce redundancy both in computer information and in natural languages. In the former, many data-compression methods have been developed to reduce redundancy in data and to save space. In the latter, commonly-used words are always short.

Perhaps the simplest way to use redundancy with digital data is the check bit and especially the parity bit, as the following example shows. Given a long bitstring, we partition it into groups of seven bits and append an eighth, parity bit to each group. The parity bit is set such that the total number of 1 bits in the 8-bit group is odd. The new bitstring is now $8/7 \approx 1.14$ times longer, but it is more reliable or robust. We can consider the added reliability as a tradeoff against the extra length, and we can philosophize and claim that there is always a price to pay for any improvement; nothing comes completely free.

Groups of eight bits are called bytes, but in general, our groups are referred to as codewords.

A single parity bit is easy to use, but it is limited to error detection (no automatic correction) and it does not provide absolute reliability. When a bitstring with parity bits is received, the receiver checks each 8-bit group by counting the number of 1-bits. If the number is odd, the receiver assumes that the group is correct, but this assumption may be wrong if an even number of bits are bad. If the number of 1-bits is even, the receiver knows that the group is bad and has an odd number of corrupt bits (i.e., one, three, five, or seven bits are bad, among them possibly the parity bit itself).

(Definition of parity. The state or condition of being equal, especially regarding status or pay. In Mathematics, the parity of a number is the fact of being even or odd. From Latin paritas, from par, meaning equal.)

◇ **Exercise 10.2:** What is the advantage of using odd parity over even parity?

Here is how redundancy creates reliability in digital data. We imagine a language with four letters `A`, `B`, `C`, and `D`, where strings consisting of these letters have to be digitized and transmitted. The simplest way to digitize four symbols is to assign them the codes 00, 01, 10, and 11, respectively. The string `ABBDAC` would then be converted to 00|01|01|11|00|10 (without the vertical bars). When such a string is transmitted and received, there is no way to detect any possible errors, because any error would change a 2-bit string to another 2-bit string, but all four possible 2-bit strings have already been assigned and are valid codewords.

A simple solution is to add a third bit to the codes and choose, say, the four strings 000, 001, 010, and 011 as the codes of our four symbols. If code 001 is damaged in transmission and is received as 101, the receiver detects an error, since 101 is a non-valid codeword. However, if 001 is received as 000, the error would go undetected. A little thinking suggests that the set of codes 111, 001, 010, and 001 is a better choice because any 1-bit change in any of these codes turns it into a non-valid code. We say that the codes of this set are more distant from one another, or that the set of codes has a larger Hamming distance, a distance of 2. Changing one of these codes to another requires changing two bits. Changing one bit in any of them changes it from a valid codeword to an invalid one, which is how the set of codes can detect all one-bit errors.

The Hamming distance of two codes is the number of bit positions where they differ. The Hamming distance of a set of codes is the maximal distance between the codes of the set. This important concept is named after Richard Hamming who in the 1950's has developed the concept of distance as a general way to use check bits for error detection and correction. In general, a set of codes with a Hamming distance of $d+1$ can detect all d-bit errors.

Symbol	code_1	code_2	code_3	code_4	code_5
A	0000	0000	001	001001	01011
B	1111	1111	010	010010	10010
C	0110	0110	100	100100	01100
D	0111	1001	111	111111	10101
m:	2	2	1	4	3

Table 10.15: $(k = 2)$.

Table 10.15 lists several error-control codes. Code_2 has a Hamming distance of 2, so it can detect all single-bit errors. Code_3 features the same Hamming distance, but is shorter. Code_4 has a Hamming distance of 4, sufficient to detect all 3-bit errors. It was constructed as two copies of code_3, which makes it obvious that codes with longer Hamming distances can easily be created. However, such codes may be too long, which is why clever, efficient approaches to code design are important.

Notation. Given a data file with symbols that are k-bits long, we can translate each symbol to an n-bit codeword by adding m bits for increased redundancy. Thus, $n = k + m$. We use the standard notation (n, k) for such a code.

The ratio n/k gives a measure of the overhead of a code. Notice that $n/k = \frac{m+k}{k} = 1 + m/k$, where m/k is the overhead of the code. In the case of code_1 the overhead is 2, and in the case of code_3 it is 3/2.

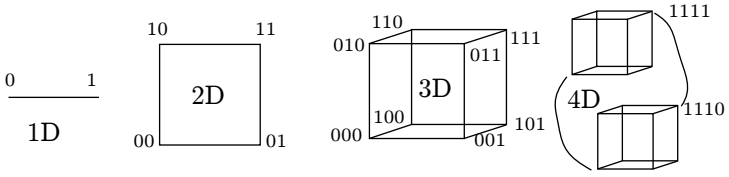


Figure 10.16: Cubes of Various Dimensions and Corner Numbering.

For mathematically-savvy readers. The definitions above have a simple geometric interpretation. Imagine a hypercube in n-dimensional space. Each of its 2^n corners can be numbered by an n-bit number (Figure 10.16) such that each of the n bits corresponds to one of the n dimensions. In such a cube, points that are directly connected (near neighbors) have a Hamming distance of 1, points with a common neighbor have a Hamming distance of 2, and so on. If a code with a Hamming distance of 2 is needed, only points that are not directly connected should be selected as valid codewords.

Once the idea of the Hamming distance and its use for error detecting is clear, its extension to error correcting is obvious, just increase the Hamming distance of the set of codewords. When the receiver identifies a bad codeword, it corrects the error by selecting the codeword that is closest to the bad one. An example is code_5 of Table 10.15, which has a Hamming distance of 3. When one bit is changed in any of its four codewords, that codeword is one bit distant from the original one but is still two bits distant from any of the other codewords. Thus, if there is only one error, the receiver can always correct it.

What code will correct 2-bit errors? When two bits go bad in a codeword A, the resulting codeword B is at a distance of 2 from A. Such an error can be corrected if the distance between B and all other codewords is greater than 2, i.e., at least 3. If this is true, then A is the codeword closest to B and the error is corrected by changing B to A. Thus, to correct 2-bit errors, a code with Hamming distance 5 or greater is needed.

In general, when d bits get corrupted in a codeword A, it turns into a codeword B at a distance d from A. (Notice that B should be invalid, which means that the code must have a Hamming distance of at least $d+1$.) If the distance between B and any other valid codeword is at least $d+1$, then B is closer to A than it is to any other valid codeword. This is why a code with a Hamming distance of $d+(d+1)=2d+1$ (or greater) can correct all d-bit errors.

How are the codewords selected? The problem is to select a good set of 2^k codewords out of the 2^n possible ones. The simplest approach is to use brute force. It is easy to write a computer program that will examine all the possible sets of 2^k codewords and select the first one that has the right distance. The problems with this approach are (1) the time and storage required at the receiving end to verify and correct the codes received, and (2) the amount of time it takes to examine all the possibilities.

Problem 1. The receiver must have a list of all the 2^n possible codewords. For each codeword, it must have a flag indicating whether the codeword is valid, and if not, which valid codeword is the one nearest it. The list has to be searched for each codeword received in order to verify it.

Problem 2. In the case of four symbols, only four codewords need be selected. For code_1 and code_2, these four codewords had to be selected from among 16 possible numbers, which can be done in $\binom{16}{4}=7280$ ways. It is possible to write a simple program that will systematically select sets of four codewords until it finds a set with the required distance. In the case of code_4, however, the four codewords had to be selected from a set of 64 numbers, which can be done in $\binom{64}{4}=635{,}376$ ways. This is still feasible, but it illustrates the magnitude of the problem. In practical cases, where sets of hundreds of symbols are involved, the number of possibilities of selecting codewords may be too large to handle in a reasonable time frame.

There are many techniques and approaches to constructing sets of error-control codes, and out of all of them, this discussion concentrates on the Reed-Solomon codes. Those codes were chosen to maintain the reliability of compact discs (CDs) when those devices were developed in the late 1970's, so we precede the discussion of the Reed-Solomon codes with a short survey of the history, properties, and structure of CDs.

> Mistakes are almost always of a sacred nature. Never try to correct them. On the contrary: rationalize them, understand them thoroughly. After that, it will be possible for you to sublimate them.
>
> —Salvador Dali.

10.5 Compact Disc (CD)

As of 2023, CDs are seldom produced or utilized. The DVDs, which followed CDs, are also gradually falling out of common usage. Computer users and music enthusiasts may find it difficult to comprehend the level of excitement generated by these devices upon their public release in 1982. In earlier times, music was kept and listened to on vinyl records and cassette tapes. These devices operated in an analog manner. Their capacities were limited, lasting only a few minutes of play, and they were delicate and prone to damage from scratches and rough handling. Additionally, replicating analog information consistently reduces its quality.

The compact disc (CD) was developed by Philips and Sony, starting in 1974. In June 1980 the two companies agreed on a common CD standard. In 1981 this standard was approved by the Digital Audio Disc committee, was ratified as standard IEC 908, and has been used since. The standard includes details of the signal format, disc material, and error-correcting code. Note the spelling "disc," as opposed to a magnetic "disk."

The original CD was designed to hold up to 74 minutes of uncompressed stereo digital audio or about 650 MB of data. Capacity is routinely extended to 80 minutes and 700 MiB by arranging data more closely on the same-sized disc. The immediate success of the audio CD motivated the development of the CD-ROM (for digital information) in 1985, the CD-V (for video), and the CD-I (interactive), both in 1987. References [Pohlman 92] and [Peek 09] have detailed information on the history and technology of this important historical product.

Physically, the CD is a disc, 1.2 millimeters thick, with a 120 mm diameter. The hole at the center is 15 mm in diameter. The distance between the inner and outer circumferences is therefore $(120 - 15)/2 = 52.5$ mm. Of this, only 35 mm is actually used, leaving safety margins both inside and outside. The information is recorded on a metallic layer (typically aluminum, silver, or gold), that is 0.5μ to 1μ thick (where μ, or micron, is 10^{-6} meter). Above this layer there is a protective lacquer coating (10μ to 30μ thick), with the printed label. Below the metal layer is the disc substrate, normally made of transparent polycarbonate. It occupies almost the entire thickness of the disc. Since the protective layer on top is so thin, any scratches on the label can directly damage the metallic layer. Even pen marks can bleed through and cause permanent damage. On the other hand, scratches on the substrate are usually handled by the error correcting code (Section 10.6).

Figure 10.17 illustrates the bits recorded on the spiral of the CD, and it is obvious that careless handling of this disc can cause bursts of errors in consecutive bits. This is why a CD should be cleaned with a radial motion, from the center to the rim, rather

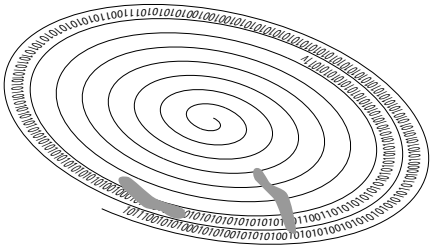


Figure 10.17: Scratches on a CD.

than a circular motion, along the spiral track. Such careful cleaning may also introduce errors, but they would tend to be scattered rather than concentrated in groups.

The digital information is recorded in pits arranged in a spiral track that runs from the inner circumference to the outer one. The pits are extremely small. Each is 0.5μ wide and 0.11μ deep (Figure 10.18b). Pit lengths range from 0.833μ to 3.56μ. The track areas between pits are called *land*. The distance between successive laps of the track is 1.6μ. As a result, the track makes 22,188 revolutions in the 35 mm recording area. Its total length is about 3.5 miles. The information is recorded such that any edge of a pit corresponds to binary 1, and uniform areas in pits and in lands (between pits) correspond to successive zeros. To reduce fabrication problems, the pits should not be too short or too long, which means that the number of binary ones recorded should be carefully controlled (see below).

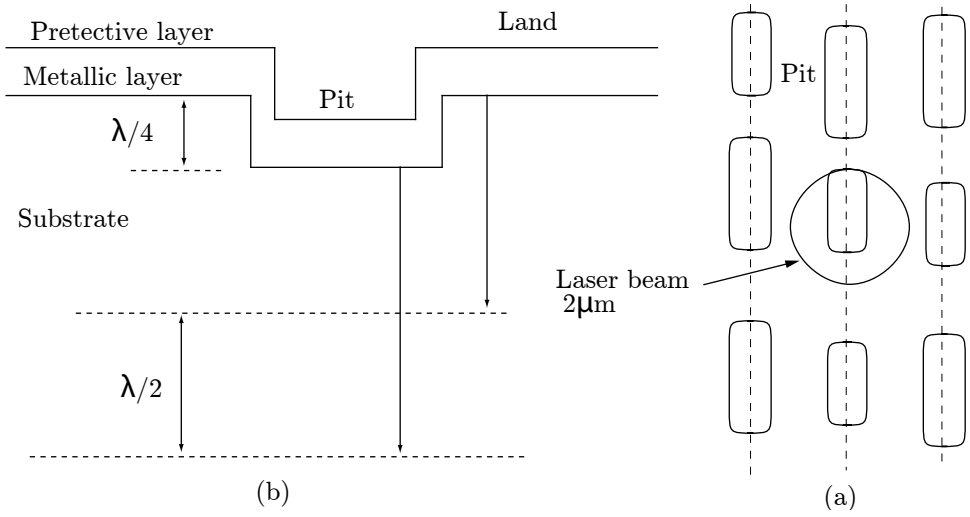


Figure 10.18: Lands and Pits on a CD.

To read the disc, a laser beam is focused on the track through the disc substrate, and its reflection is measured (Figure 10.18a). When the beam enters a pit, the reflection drops to virtually zero, because of interference. When it leaves the pit, the reflection goes back to high intensity. Each change in the reflection is read as binary one. To read the zeros, the length of a pit, and the length of a land between pits, must be measured accurately.

10.6 Reed–Solomon Codes

This section presents an easy to understand, step by step geometric approach to the popular Reed–Solomon (RS) codes. These codes are commonly used for error correction in a wide range of applications such as CDs, DVDs, QR codes, barcodes, the voyager spacecrafts, and satellite communications. To the many books, articles, and videos that explain and describe these codes, I would like to add [vivint 17] and [tomverbeure 22].

It is obvious that reading the digital data of a CD-ROM must be error free, but error correction is also important in an audio CD, because one bad bit can cause a big difference in the note played. Consider the two 16-bit numbers 0000000000000000 and 1000000000000000. They differ by one bit only, yet when played as sound samples, the first corresponds to silence and the second produces a loud sound. The size of a typical dust particle is 40μm, enough to cover more than 20 laps of the track and cause several bursts of errors (Figure 10.19a). Without extensive error correction, the music may sound like one long scratch.

Any error correction method used in a CD must be very sophisticated, since the errors may come in bursts or may be individual. The use of parity bits makes it possible to correct individual errors but not a burst of consecutive errors. This is why *interleaving* is used in addition to parity bits. The principle of interleaving is to rearrange the samples before recording them on the CD and to reconstruct them after they have been read. This way, a burst of errors during reading is translated to individual errors (Figure 10.19b), that can then be corrected by their parity bits.

The actual RS code used in CDs is called the Cross-Interleaved Reed–Solomon Code (CIRC). It was developed by Irving Reed and Gustave Solomon at MIT Lincoln Lab in 1960 [Reed and Solomon 60] and is a powerful code. One version of this code can correct up to 4000 consecutive bit errors, which means that even a scratch that covers three millimeters of track can be tolerated on a CD and would automatically be corrected. The principle of CIRC is to construct a geometric pattern that is so familiar that it can be reconstructed even if large parts of it are missing. It's like being able to recognize the shape of a rectangular chunk of cheese after a mouse has nibbled away large parts of it.

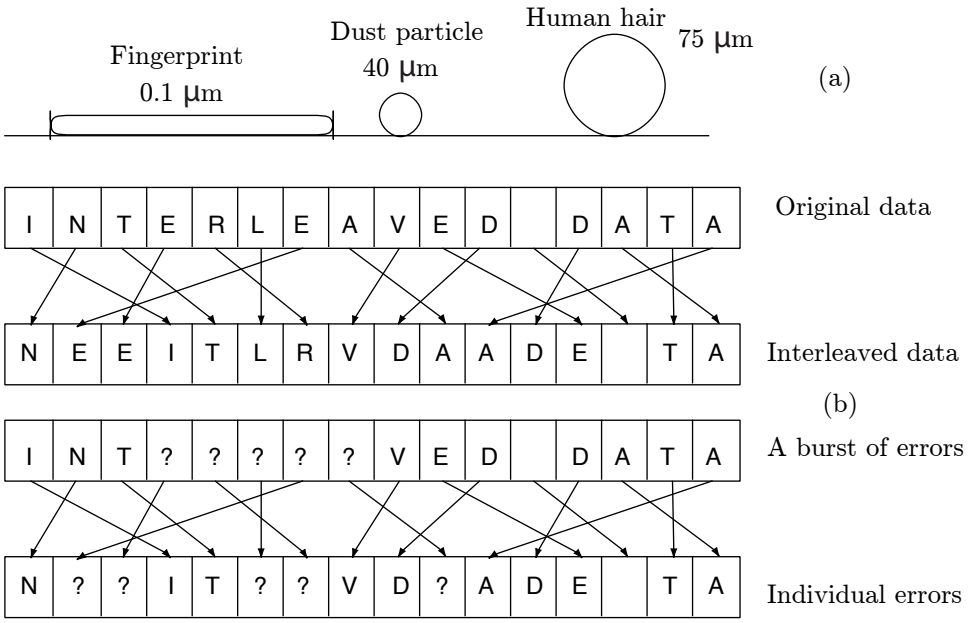


Figure 10.19: (a) Relative Sizes. (b) Interleaving Data.

Mathematician Gustave Solomon died on January 31 [1996] in Los Angeles. He was 65. Dr. Solomon was a co-inventor with Irving S. Reed of the Reed–Solomon codes, which have come into increasingly widespread use as a way of combating the inevitable errors that occur in the transmission and storage of information. He did early work concerning the algebraic theory of error-correcting codes, and, with H. F. Mattson, was co-author of the powerful tool for analyzing such codes known as the Mattson-Solomon polynomial. His other interests included composing popular songs and folksongs and teaching voice and movement.

—Dave Farber, `farber@central.cis.upenn.edu`

The RS code is introduced here in steps, using geometric figures and their relations to polynomials. The first step is based on straight lines; it is easy to understand, but is very inefficient. In this simple version of the RS code, the input data is read in pairs of numbers. Assume that a pair consists of the numbers 3.6 and 5.9. We consider them the y coordinates of two two-dimensional points, and we assign them x coordinates of 1 and 2, respectively. We end up with the points $(1, 3.6)$ and $(2, 5.9)$ that determine a straight line. We now compute four more points on this line, with x coordinates of 3, 4, 5, and 6. They are $(3, 8.2)$, $(4, 10.5)$, $(5, 12.8)$, and $(6, 15.1)$. Adding such points to a line or to any geometric figure is referred to as oversampling the figure. Since the x coordinates are so regular, we only need to output the y coordinates of these points. We accordingly store (or write on the CD) the six numbers 3.6, 5.9, 8.2, 10.5, 12.8, and 15.1.

Now, consider the decoding done on the receiving end. Suppose that two of those six

numbers have been damaged in cleaning or improper storage of the CD. When the new sequence of six numbers is input and checked for the straight line property, the remaining four numbers can be identified as being collinear and can still be used to reconstruct the line. Once the line is known, the two bad numbers can be corrected, since their x coordinates are known. Even three bad numbers out of the six can be corrected, since the remaining three numbers would still be enough to identify the original straight line. (Notice that there is a small but finite chance that any other set of three of the six numbers would be on a straight line, so the RS code is not perfect and may sometimes fail to correct two- or three-number errors.)

This version of the RS code is simple, but it requires four extra numbers to correct two numbers. The overhead is $4/2 = 2$, too big. The next step uses the same technique, but reduces the overhead by employing a curve, specifically a parabola, instead of a straight line.

In this step, the input data is read in blocks of three numbers. An input triplet a, b, and c is converted to the points $(1, a)$, $(2, b)$, and $(3, c)$, and the (unique) parabola that passes through them is computed. Four more points, with x coordinates of 4, 5, 6, and 7, are then computed on this parabola. Once the seven points are known, they provide a strong pattern. If even three of the seven get corrupted, the remaining four can be used to reconstruct the parabola and correct the three bad ones. However, if four of the seven get corrupted, then no four numbers will be on a parabola (and any group of three will define a different parabola). Such a code can correct three errors in a group of seven numbers, but it requires high redundancy (seven numbers instead of four). The overhead of this version of the code is $4/3 = 1.33$, lower than in the previous step, but still too high.

◇ **Exercise 10.3:** Show how to compute the parabola that passes through three given points.

Both a straight line and a parabola are special cases of polynomials, so the next step in presenting the RS code extends the technique of the previous steps to general polynomials. A polynomial $p(x)$ of degree n is the simple function

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = \sum_{i=0}^{n} a_ix^i.$$

It depends on $n+1$ coefficients a_i that are normally real numbers but can be any other mathematical entities. A degree-1 polynomial has the form $a_0 + a_1x$. When plotted, its curve is a straight line, so it can intercept the x axis at most once (except for the special, degenerate cases where the line is the x axis itself or runs parallel to it). A degree-2 polynomial has the form $a_0 + a_1x + a_2x^2$, which when plotted becomes a parabola; it intercepts the x axis at most twice (again, there are special cases). In general, the curve of a degree-n polynomial is wavy. It oscillates, goes up and down, and intercepts the x axis at most n times.

An (n, k) Reed–Solomon code, normally denoted by RS(n, k), encodes a block of k data symbols a_0, a_1,$\ldots$,a_{k-1} into a redundant codeword of n symbols, where n must be greater than k. Specifically, if n equals $k + 2t$ for a certain t, then the code can correct errors in up to t symbols (i.e., its Hamming distance is $2t + 1$). If a symbol consists of

s bits, then the numeric value of a symbol is an integer in the interval $[0, 2^s - 1]$, and there can be up to 2^s distinct symbols.

A common example is the $(255, 223)$ code with $s = 8$. Each symbol is an s-bit number (a byte), and 223 data bytes are combined with 32 extra bytes to produce a 255-byte codeword. Adding 32 bytes implies $t = 16$ (because $n = k + 2t$), so this code can correct errors in up to 16 bytes. One extreme case is where each corrupted byte has just one bad bit. In this case, the number of bits corrected is 16. Another extreme case is where each of the 16 bad bytes has been completely corrupted; every bit is bad. In this case, correcting 16 bytes amounts to correcting $16 \times 8 = 128$ bits.

The codes may be shortened in cases where no convenient value of k is available. If for some reason it is inconvenient to use $k = 223$, it is possible to use another value such as $k = 168$. The encoder appends 55 zero bytes to a block of 168 data bytes, to end up with 223 bytes. An additional 32 bytes are then computed from the 223 bytes, and a codeword consisting of the original 168 data bytes and the extra 32 bytes is sent to the decoder, where the 55 zero bytes are implied.

We describe two intuitive approaches to generating the code. In both approaches the original k data symbols a_i are considered the coefficients of the degree-$(k-1)$ polynomial

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1}.$$

- In the first approach, the polynomial is evaluated at n distinct, nonzero points x_i, and the n values $y_0 = p(x_0)$, $y_1 = p(x_1), \ldots, y_{n-1} = p(x_{n-1})$ become the codeword. The x_i values are built into both encoder and decoder. The decoder receives the codeword with the y_i's and combines each x_i with the corresponding y_i to end up with the n points (x_0, y_0) through (x_{n-1}, y_{n-1}). The points are used to reconstruct polynomial $p(x)$, whose coefficients a_i are the original data symbols. Without any errors, reconstruction is easy. The polynomial is of degree $k - 1$, so it can be determined by k points. With no errors, any k of the n points can be used, and any set of k points will determine polynomial $p(x)$.

In the presence of errors, decoding is more complex. The decoder generates the n points. It knows that the unknown, degree-$(k - 1)$ polynomial $p(x)$ passes through the points. Assume that some of the points (at most t) may be bad. A set of k points is enough to determine a unique degree-$(k - 1)$ polynomial, but given more points, this polynomial may not pass through the extra points. The decoder starts by finding a subset of at least $k + t$ points through which a degree-$(k - 1)$ polynomial $q(x)$ passes. Such a subset is termed *consistent* and it always exists, because polynomial $p(x)$ that was computed by the encoder passes through all n points, and we assume that at most t points got corrupted. (If such a subset cannot be found, the decoder concludes that more than t errors have occurred.)

Finding such a subset and determining polynomial $q(x)$ is only the first step. Some of the points in the subset may be bad, so polynomial $q(x)$ determined by the subset may be different from $p(x)$. However, the subset consists of $k + t$ (or more) points, and of those, not more than t are bad. We therefore start with the case where there are no bad points in the subset. All the points in the subset are among the ones originally used by the encoder, so polynomial $q(x)$ determined by the subset is the one computed by the encoder. In such a case, every set of k points in our subset will determine the

same polynomial $q(x)$, and this will be identical to $p(x)$. If the subset includes some bad points, then different sets of k points in the subset will determine different degree-$(k-1)$ polynomials.

In the second step, the decoder has to test every set of k points in the subset. If different sets determine different degree-$(k-1)$ polynomials, then the subset contains some bad points, and another subset must be identified and tested. If no more subsets can be found, the decoder concludes that more than t points are bad and the error cannot be corrected.

This approach is impractical because of the many calculations required and because the numbers involves are non-integers. Extensive simplifications are required in order to make the RS codes practical, so they can be used in real devices that operate in real time.

■ The next approach, while also computationally infeasible, has the advantage of being simple in the common case where no errors are discovered. The idea is to generate a codeword where the first k symbols are simply the data symbols a_i, and these are followed by $2t$ parity check symbols. The parity symbols are of the form $p(x_i)$, where the x_i are distinct, nonzero values. Such a Reed–Solomon code is known as systematic, because in the common case where there are no errors, the decoder simply outputs the first k symbols of the codeword.

To check for errors, the decoder generates the system of $2t$ equations

$$\begin{aligned} a_0 + a_1x_0 + \cdots + a_{k-1}x_0^{k-1} &= y_0,\\ a_0 + a_1x_1 + \cdots + a_{k-1}x_1^{k-1} &= y_1,\\ &\vdots\\ a_0 + a_1x_{2t-1} + \cdots + a_{k-1}x_{2t-1}^{k-1} &= y_{2t-1}. \end{aligned}$$

As with the first approach, there must be at least one consistent subset of $k+t$ symbols, and any such subset has k correct symbols. Assume that a subset with k correct symbols has $k-l$ data symbols a_i and l polynomial values $p(x_i)$. Each polynomial value $p(x_i)$ contributes an equation, for a total of l equations. Each equation has k symbols a_i on the left side, but only l are unknown, so the system of l equations can be solved.

The next steps in the description of the RS codes consider ideas for optimization. The first question is how to implement the codes using just integers. In a computer, non-integers are normally represented as floating-point numbers. Such numbers can be very small (close to zero) and very large, but they have limited precision. In addition, arithmetic operations on floating-point numbers are slow. All this complicates the problem of reconstructing the parabola or any other geometric figures. (The term limited precision means that a number such as, say, 1.3744492 can only be represented approximately, perhaps as 1.37445.)

The Reed–Solomon codes that are used in practice employ integers, because integers can be stored in the computer in full precision and operated on without loss of precision. The problem is that operations on integers may result in very large integers, in negative integers, or in nonintegers, which is why practical RS codes use arithmetic operations in finite fields. A finite field (also known as Galois field, GF) is a finite set of elements

(integers or other symbols) with two special arithmetic operations defined on them, operations that always result in a field element (the field is closed under each operation). Here is a short summary for those familiar with finite fields.

An (n, k) Reed–Solomon code deals with elements of the finite field GF$(n+1)$, which implies that $n+1$ should be either a prime or a power of a prime. This code encodes a group of k field elements into a group of n field elements. A common example is the RS $(255, 223)$ code that is based on the field GF(256). Each element of this field is a byte, and the code encodes a group of 223 field elements into 255 elements by adding 32 check field elements. It can be shown that an (n, k) RS code has a Hamming distance of $2t+1$, where t is defined by $2t = n - k$. The code can therefore correct up to t errors. (The following is another interpretation of these numbers. Start with k data items, normally bits, and increase the data redundancy by adding $2t$ data items. The resulting code is $n = k + 2t$ long, and it can correct up to t errors.)

Given a finite field GF$(n + 1)$ and a string $a_0a_1 \dots a_{k-1}$ of k field elements to be encoded, the (n, k) RS code is computed in the following steps:

1. Select a primitive element g of the field. This is a generator of the multiplicative cyclic group of the nonzero elements of the field.
2. Treat the message as the polynomial $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{k-1}x^{k-1}$.
3. Compute this polynomial at the n points $p(1)$, $p(g)$, $p(g^2)$,..., $p(g^{n-1})$.

The n field elements computed in step 3 are the RS code of the original string. This code can be stored or transmitted instead of the string. Because of the way g is selected, the various powers g^i are nonzero and are all different. Returning to the geometric interpretation, these n elements are n points on the graph of the polynomial $p(x)$. This polynomial is of degree $k - 1$, so k points are enough to determine it uniquely. The polynomial is computed at n points where $n > k$, so it is overcomputed by $2t = n - k$ points. Because of the way this particular polynomial is computed, it has the following useful property: Given any k of the n points, if we compute the degree-$(k-1)$ polynomial that passes through them, the result will be $p(x)$.

Finally, the last and ultimate optimization is to implement the RS decoder in hardware. The digital computer is a very useful machine because it is general purpose; it can perform many tasks by running different software. However, specialized hardware is always the fastest way to perform a computation that requires speed. Thus, one of the main innovations in planning and building CDs (and in other applications of the RS codes) was the fabrication of special software to implement the decoding part (encoding is done once, before a batch of CDs is manufactured, but decoding is done each time the CD or DVD is played, and so must be performed in the small player unit and in real time.

Decoding is complex and is beyond the scope of this book. It involves five main steps: (1) syndrome calculation, (2) Berlekamp algorithm, (3) Chien search, (4) Forney's algorithm, and (5) error correction. For one of several detailed references that describe this process, see [Berlekamp 68].

Another advantage of the RS code is its excellent behavior for burst errors. Its Hamming distance is $2t + 1$, so it can correct up to t errors, but if the locations of the errors are known, all of them (up to $2t$ errors) can be corrected. This process is easy to visualize when we consider the geometric interpretation. If we know the bad points,

we can reconstruct the polynomial from k points, then correct up to $2t$ bad points by "pulling" them into the polynomial.

From the errors of others, a wise man corrects his own.
—Publilius Syrus.

10.7 SVD Image Compression

Someone who constantly objects to anything he hears or is told can be called a contrarian or argumentative. Such a person likes to object, to start an argument, and to always claim superior knowledge. However, there is one phrase to which even the most extreme contrarian cannot object, and this is "an image is worth a thousand words." It seems that images have two attributes that make them vital to us; they convey much information while also able to entertain us.

It has long been known that inventions and developments that give us access to images tend to be successful. Photography, movies, television, and YouTube are only a few examples. We use images, we are used to having images around us and in front of us. We keep images, mostly in the form of photographs, and we loathe to part with them, even knowing that the chance we would ever look at most of our older images is slim.

However, images have one important shortcoming, they tend to be large. This is no wonder considering the amount of information in an image when compared to text. This is why the compression of images is an important area of computer science, and why many different approaches, algorithms, and implementations have been developed over the years to compress images efficiently and decompress them fast.

Most image compression methods are based on the discrete cosine transform or on wavelets. Those methods are powerful and have been discussed in many texts and videos. The method described here is mathematically simpler and much less known. It is based on a technique known as singular value decomposition (SVD) of matrices.

Note. This section assumes basic knowledge of matrices and their operations. Only two matrix-related concepts are defined here:

- A unitary matrix is one whose inverse equals its transpose.

- The rank of a matrix A is the maximal number of linearly independent columns of A, a number which always equals the maximal number of linearly independent rows of A. The following matrix has rank 1, because any of its columns c_i can be generated as a linear combination $\alpha c_j + \beta c_k$ of other columns (and the same is true for its rows)

$$A = \begin{pmatrix} 1 & -1 & 2 \\ 2 & -2 & 4 \\ -4 & 4 & -8 \end{pmatrix}.$$

An arbitrary $n \times n$ matrix consists of n^2 elements, but a rank-1 matrix can easily be represented with fewer elements. Matrix A above can be written as the inner product

$$A = \begin{bmatrix} 1 \\ 2 \\ -4 \end{bmatrix} [1, -1, 2], \tag{10.2}$$

where this notation means (1) multiply row $[1, -1, 2]$ by 1 to become the top row of A, (2) multiply the same row by 2 to become the second row of A, and multiply it by -4 to become the third row. In this way we can represent the nine elements of A by six numbers, and achieve a slight compression. This idea becomes more interesting when dealing with large matrices, such as the pixels of a typical digital image. Such an image, which currently consists of at least several megapixels, can be compressed significantly in this way. There is only one problem, the pixel matrix of a typical image is of a high rank. Its columns tend to be independent of one another.

The next idea is to start with an arbitrary matrix and decompose it into a sum of rank-1 matrices which can then be compressed as shown above. The efficiency of this approach depends heavily on the length of the decomposition. If the decomposition of a 3×3 matrix consists of three rank-1 matrices, then it ends up expanding the original $3 \times 3 = 9$ elements to $3 \times 6 = 18$. Thus, we need a special decomposition algorithm that results in a sum of inner products with smaller and smaller elements. In such a case, several inner products at the end of this sum can be deleted, which may result in significant compression.

Naturally, the price to pay for such an elegant but lossy algorithm is loss of image detail, but we know from long experience that images can suffer a certain amount of loss when compressed, and still appear un-degraded to an average viewer when decompressed.

A matrix can be decomposed into a linear combination of other matrices in different ways, but there is one decomposition, namely the Singular Value Decomposition (SVD), which is optimal in the sense that it has many practical applications, one of which is the (lossy) compression of digital images. (Being lossy implies that it is limited to images and cannot be used for text compression.) The basic idea behind SVD is that any matrix M can be decomposed to become the product $U \cdot \Sigma \cdot V$ of three matrices, and can be reconstructed by multiplying $U \cdot \Sigma \cdot V^T$. For simplicity, we assume that our matrices are square, $n \times n$, but SVD can also handle rectangular matrices as well as complex numbers. Thus, for example, the Mathematica statement

```
{u, s, v} = SingularValueDecomposition[{{1.,-2,6},{2,4,8},{-3,18,7}}]
```

returns the three $n \times n$ matrices (shown here with a precision of only two digits)

$$U = \begin{bmatrix} -0.04 & 0.70 & 0.70 \\ -0.34 & 0.65 & -0.6 \\ -0.93 & -0.27 & 0.21 \end{bmatrix} \Sigma = \begin{bmatrix} 20.64 & 0 & 0 \\ 0 & 8.87 & 0 \\ 0 & 0 & 1.35 \end{bmatrix} V = \begin{bmatrix} 0.10 & 0.31 & -0.94 \\ -0.88 & -0.41 & -0.23 \\ -0.46 & 0.85 & 0.23 \end{bmatrix}$$

while executing `u.s.Transpose[v]` reproduces the original matrix (up to small numerical errors caused by limited machine precision).

What makes SVD useful for our purposes are the following special properties:

- U is unitary and therefore also orthogonal.

- Σ is diagonal.
- V is orthogonal.
- The SVD decomposition is not unique. It can always be performed such that the elements σ_i of Σ, which are on its diagonal, are in descending order from top-left to bottom-right.

As a result of these properties, it is possible to write the SVD decomposition of M as the sum

$$M = \sum_{i=1}^{n} \sigma_i u_i v_i^T,$$

where u_i is column i of U and v_i^T is row i of V. Even more, because elements σ_i are in descending order, some of the last ones may be zero, which reduces the sum to fewer than n terms. Using the inner product notation of Equation (10.2), here are the details of the first two terms of this sum:

$$\sigma_1 \begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{bmatrix} [v_{11}, v_{12}, \ldots, v_{1n}] + \sigma_2 \begin{bmatrix} u_{12} \\ u_{22} \\ \vdots \\ u_{n2} \end{bmatrix} [v_{21}, v_{22}, \ldots, v_{2n}] + \cdots. \tag{10.3}$$

Thus, even though matrices U and V are not of rank 1, each term of (10.3) is small and consists of a single σ_i number, a column, and a row, for a total of $1 + 2n$ numbers.

Our interest is in compression, so we count the number of matrix elements so far. Each term of (10.3) consists of $1+2n$ numbers, so the complete sum of n terms has a total of $n(1+2n) = n+2n^2$ numbers, more than twice the $n \times n$ numbers of the original matrix. It seems that in our wisdom we have managed to end up with significant expansion, but in practice we find that it is easy to achieve high compression rates and a high-quality decompressed image by deleting many terms from the end of this sum.

Recall that elements σ_i are getting smaller with increasing i, causing successive terms of (10.3) to become smaller and smaller. This, combined with the fact that images have much image noise, makes it possible to retain only the first few terms and still be able to reconstruct the original image without any visible degradation. A typical example is a 16 Mpixel photograph, where $n = 4,096$. The (10.3) sum of such an image can typically be cut to just a few dozen terms, resulting in a high compression ratio.

This discussion describes the principle of image compression with SVD. Any real implementation has to deal with practical problems, and the following point is especially crucial. The number of terms that should be kept depends on the image that is being compressed. Images with much image noise (i.e., where nearby pixels tend to be different from one another) may require more terms, resulting in poorer lossy compression. Any implementation must therefore be user friendly. It should let the user interact with the algorithm in real time, and specify the number of terms to keep experimentally.

10.8 What is Average?

A typical dictionary definition of the common term "average" is: An average is a single number or value that best represents a set of data. For most people, the average of numeric data is the arithmetic mean, the sum of the numbers divided by the amount of numbers in the data to be averaged. However, mathematicians, being fussy and meticulous, distinguish several types of numeric averages, which are mentioned here, but without examples.

- Arithmetic mean (AM or arithmetic average) of the numeric list a_1, a_2, $\ldots a_n$, is

$$\frac{1}{n}\sum_i a_i.$$

- The geometric mean, GM, is the nth root of the product of n numbers. Thus, $\sqrt[n]{a_1 \times a_2 \times \ldots \times a_n}$. This type of average is often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as a set of growth figures, values of a population, or interest rates of loans or investments. The Khinchin constant (Page 62) employs this type of average.

- The harmonic mean, HM, is the reciprocal of the arithmetic mean of the reciprocals of the items being averaged, which should be positive real numbers. This type of average is sometimes appropriate for situations when the average rate is desired. The mathematical expression is

$$H(a_1, a_2, \ldots, a_n) = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}.$$

The harmonic mean is based on reciprocals, which explains why it amplifies the effect of small data items and mitigates the effect of large items. It is therefore often used whenever it is desired to reduce the impact of large outliers and magnify the impact of small ones.

The three types of average above are also known as the Pythagorean means. They were recognized and studied by Greek mathematicians because of their applications to geometry.

- The quadratic mean (QM or root mean square, RMS), of a set of values is the square root of the arithmetic mean of the squares of the values. Thus

$$x_{\mathrm{RMS}} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)}.$$

The following relations exist between the three Pythagorean means. Given a data set of positive reals that are not all equal, it is true that QM > AM > GM > HM. If all data items in the set are equal, the three Pythagorean means are also equal.

There are also weighted averages, where a numeric weight is assigned to each data item. Such an average is useful when it is known that certain items are more important than others.

There are also measures of central tendency, of which the following are the chief ones:

- The mid-range (or mid-extreme) of a set of data items x is the arithmetic mean of the maximum and minimum values of the data set. Thus, $(\max x + \min x)/2$.

- The median of a data set is the value separating the higher half from the lower half of the set. It may also be considered "the middle" value. If the set consists of an odd number of values, there is a middle item. Otherwise, the two middle items are arithmetically averaged to produce the median of the set.

- The mode is the value that appears most often in a set of data values. The mode of a data set may not be unique, as for example, in $(1, 1, 2, 3, 3)$, which has two modes.

For many other types of average and mean, see Wikipedia article "Average."

10.9 The power of the XOR

The Prisoner is a 1967 British television series about the abduction of an unnamed British intelligence agent. After suddenly and angrily resigning his job, he unexpectedly finds himself a prisoner held in comfort in a quiet, lovely rural place that is referred to only as *The Village*. Everyone in the village has a number instead of a name, and our hero, due to his past rank, receives the number six, to which he repeatedly refuses to answer, claiming "I will not be pushed, filed, stamped, indexed, briefed, debriefed, or numbered!"

The elegant problem presented here, following the video at [almost 20], is best described in terms of two prisoners. They are offered their freedom if they can win the following mathematical game against their warden. The warden shows them an 8×8 board on which are written 64 random bits, and explains the rules of the game he has in mind. (We assume that it is easy to distinguish the top and bottom of the board.) The prisoners are then given time to discuss their freedom-winning strategy. Prisoner #2 then leaves the room, and the warden, in the presence of prisoner #1, hides a secret key under one of the 64 squares of the board. Prisoner #1 then leaves a hint on the board, to help #2 find the key. The hint is to flip one of the 64 random bits on the board. Prisoner #1 then leaves and #2 is brought in. He cannot identify the flipped bit, but if he employs the correct strategy to locate the key, he can gain freedom for both himself and his friend.

A trivial strategy may be for #1 to flip the bit over the key, but this won't work because #2 cannot tell which bit had been flipped.

At first, this seems impossible. How can the unknown location and value of one flipped bit lead to the location of another bit among 64 similar-looking ones? But this game can be won due to the power of an important logical operation known as exclusive-OR, XOR, or $\oplus$ for short. We know that computers use binary (base-2) numbers. We also know that there is nothing special about 10, and that any integer greater than 1 can be used as a number base. See also Section 2.2.

However, there is something special about 2 as the base of a numbering system, because the use of binary numbers allows for an entirely new class of *logical* operations,

in addition to the familiar arithmetic operations. Of special interest is the exclusive-OR logical operation, whose definition is extremely simple. The XOR of two different bits is 1 and the XOR of two identical bits is 0. Alternatively, the XOR is the same as adding two bits modulo 2. Given two n-bit numbers A and B, we compute the XOR of each pair of corresponding bits of A and B to end up with the n-bit $A \oplus B$. The special property of XOR that makes it very useful in many applications (see, for example, reference [XOR 23]) is its associativity. If $C = A \oplus B$, then $A = B \oplus C$.

Figure 10.20 illustrates the approach to the solution of this intriguing problem. The figure shows a smaller, 4×4 board, but the same technique works for bigger boards. The idea is for prisoner #1 to employ the binary pattern on the board—whether random, non-random, or even uniform—to first compute a number P (a parity) whose size can serve as a location on the board. For a 4×4 board (16 squares), P should be a 4-bit number, while for a 8×8 board (64 squares), P should be a 6-bit number.

Once Prisoner #1 gets the location K of the secret key from the warden, he makes use of the XOR operation to determine the location F of the bit to be flipped, such that the new parity of the board would be K. It all depends on how the parity P of the board is computed. If it is computed cleverly, then the single XOR operation $F = P \oplus K$, would give the location F of the bit to be flipped. Once this bit is flipped by #1, it changes the parity of the board from the original P to K.

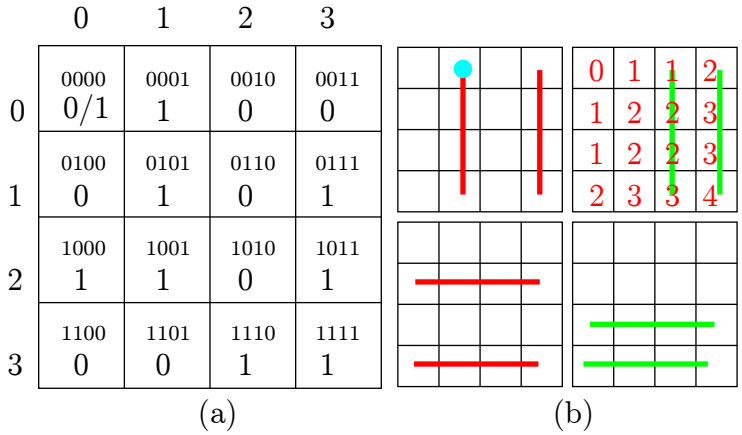


Figure 10.20: Parities for a 4×4 Random Grid.

Prisoner #1 now leaves the room, #2 comes in, computes the parity P of the board (without knowing which bit had been flipped), and looks for the key at the location pointed to by P. Success; followed by freedom.

Here is an example. Suppose that $P = 4 = 0100_2$ and $K = 11 = 1011_2$. The operation $0100 \oplus 1011$ yields $F = 1111 = 15$ (these are the bit positions where P and K differ). We have to compute the parity P in such a way that the step of flipping the bit at location 15 would change the parity of the entire board from P to K. This is the essence of our problem.

First, a few words about parity. The (odd) parity of a set of bits is 1 if the number of 1's in the set is odd, and is 0 otherwise. An even parity is the opposite.

Our strategy is to partition the board into four overlapping parts (six in the case of an 8×8 board) and compute the odd parity of each. The four (six) bits computed in this way would constitute our parity P. Figure 10.20a shows how each board location is assigned a 4-bit binary address. Once this is done, board locations are scanned and parity bits are determined in four steps as follows:

Step 1. Scan all the locations whose least-significant address bits are 1. Those are the two vertical red columns in Figure 10.20b. The total number of 1's in those columns is 6 (even) in the figure, so we prepare a parity bit of 0 and it becomes the least-significant bit of P.

Step 2. Scan the board locations whose second least-significant address bits are 1 (the two vertical green columns). There are four 1's, so the second least-significant parity bit of P is also 0.

Steps 3 and 4 are the same for the rows (the horizontal bars in the figure), prepending the two parity bits 10 to P. Thus, $P = 10\,00$.

Notice that the top-left location of the board is not scanned in this process. Its value is not used in computing P. This will be discussed shortly.

Now imagine that the key is hidden in location $K = 9 = 1001_2$. The operation $P \oplus K = 1000 \oplus 1001$ produces $F = 0001 = 1$. Prisoner #1 now flips the bit at location 1. The figure shows that this location was used in the parity computation in step 1 (the blue circle in the figure), so it has flipped bit 1 (second from the right) of the parity of the entire board. This parity, which originally was $P = 1000$ is now $1001 = 9$. Once Prisoner #2 computes the parity of the board that he sees, he points to location 9 and smiles; freedom is sweet.

Notice that this technique works even if the binary pattern of the board is non-random or even uniform, resulting in either $P = 1111$ or $P = 0000$, where the XOR operation is trivial. Having the pattern in the first place is a red herring, intended to confuse the prisoners.

Finally, here is a logical, non-mathematical explanation of why the group of squares have been chosen by their address bits. A careful scrutiny of the figure shows that each square belongs to a number of groups as follows (see the 16 red numbers in Figure 10.20b):

Square 0 belongs to zero groups. In the figure, its value is therefore undetermined.
Squares 1, 2, 4, 8, belong to one group each.
Squares 3, 5, 6, 9, 10, 12, belong to two groups each.
Squares 7, 11, 13, 14, belong to three groups each.
Square 15 belongs to four groups.

Also, each group can be identified by one bit that's common to the addresses of all the squares in the group. Therefore, any combination C of groups has at least one square that is a member of all the groups in C.

We can now modify the 4-bit parity of the board in any possible way by flipping exactly one bit. For example, if we want to change bits 0 and 2 (least-significant bit and 3rd bit from the right) of the parity, we perform the following steps:

1. Find the group whose members are the squares with a least-significant address bits of 1.

2. Find the group whose members are the squares with a 3rd-from-the-right address bits of 1.

3. Flip the bit on the square that is a member of these two groups.

A final note. Suppose that the parity of the board turns out to be $1100_2 = 12$, and the key happens to be hidden in square 12. In this case there is no need to modify the parity, and a careful check will verify that the three steps above would end up selecting Square 0, the one that belongs to no group.

See also reference [cantor.paradise 20] for other approaches to this problem.

10.9.1 Vernam Cipher

The Vernam cipher (or Vernam code) is an important example of the power of XOR. The cipher has two inputs, a plaintext and a (random) key. In the simplest case, both are bitstreams. The encoder reads the first bit from the plaintext and XORs it with the first bit from the keystream to produce the first bit of the ciphertext. This simple step is repeated for subsequent bits. The code has two important advantages: (1) It is symmetric. The Vernam decoder inputs a bit from the ciphertext and XORs it with a bit from the keystream (the same keystream used by the encoder) to produce a bit of plaintext. The encoder and decoder may therefore be the same machine. (2) The ciphertext output by the Vernam encoder is random. It has no patterns that a codebreaker can latch on to and use as guesses to break the code. This is true even if the plaintext isn't random.

This interesting and important cipher was developed by Gilbert Vernam in 1917, with United States patent #1310719 issued to him in 1918.

◇ **Exercise 10.4:** Show that if the keystream of a Vernam cipher is a random sequence of bits, then the ciphertext is random even if the plaintext isn't random.

Naturally, the three streams, plaintext, keystream, and ciphertext don't have to be bit strings. Each may consist of fixed-length records, such as bytes.

◇ **Exercise 10.5:** Explain why the following claim makes sense. The chance that a wrong key during Vernam decoding would produce meaningful but wrong plaintext is very small.

10.10 Brouwer fixed-point theorem

The interesting and unusual Brouwer fixed-point theorem is due to the Dutch mathematician L. E. J. Brouwer who is known, among other things, for the Brouwer Hilbert controversy. In its formal form, the theorem states that for any continuous function f mapping a nonempty compact convex set to itself, there is a point x_0 such that $f(x_0) = x_0$. Here is what it means, in plain words.

> O, rocks, tell us in plain words.
> —Molly Bloom in James Joyce's *Ulysses*, 1922.

Imagine a set of points which is bounded and has no holes. If the points of the set are transformed continuously, without cutting or gluing them, then at any point

during the transformation, there will be at least one point still at its original spot. The fixed point may vary over time during the transformation, but the general claim of the theorem is often surprising and is certainly unexpected.

If the set has no bounds, or has holes, it is possible for every point to change its position after a transformation. The same is true if a part of the set is cut and is pasted elsewhere in the set.

This is true regardless of the dimension of the set. Figure 10.21 (see also Figure 1.23) is a two-dimensional example. A map is copied, shrunk, rotated and translated. The fixed point is indicated by the red arrow. An example of a three-dimensional point set is the set of molecules of a liquid in a container. As long as there are no spills, the set of molecules is bounded. When we stir the liquid, it seems that *all* the molecules must move and change positions randomly, but the theorem claims that at any point in time, there is at least one molecule that happens to be at its original location. It is like saying that the process of stirring a liquid must always remain imperfect. (This claim assumes that molecules are small enough to closely resemble mathematical points.)

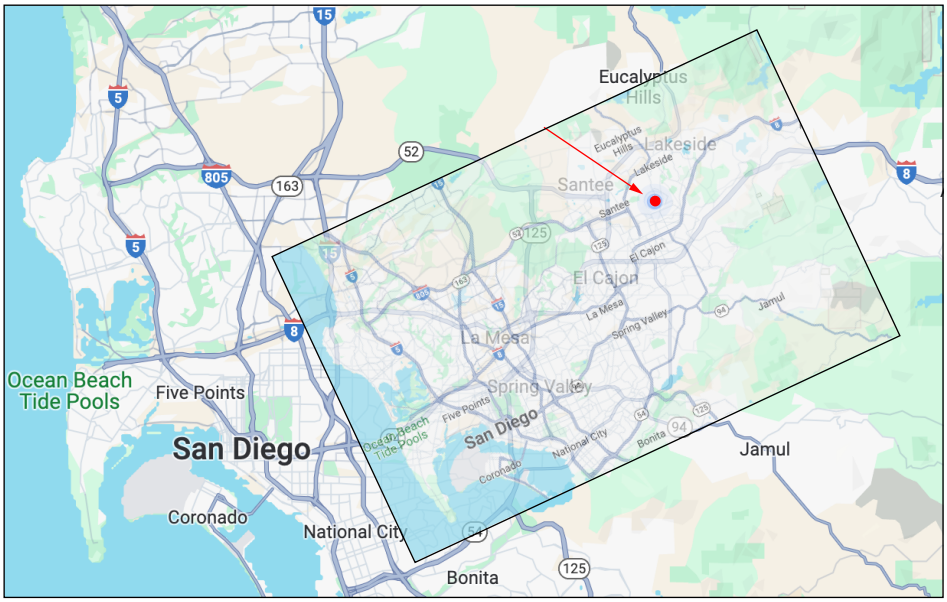


Figure 10.21: A Fixed Point in Two Dimensions.

A related result is the Borsuk-Ulam theorem. At any point in time, there is at least one pair of points on the Earth's surface that are diametrically opposed and have the same temperature and atmospheric pressure. Diametrically opposed points on a sphere are called antipodes (pronounced antipodés).

The proof of this surprising result is simple. We start with temperature. Choose a point A on the surface of the earth. Denote the diametrically opposed point by B. If the two are at the same temperature, we are done. Otherwise, let's assume that A's temperature is higher than B's. We now move them along an arbitrary path in order to

eventually swap their locations. They can be moved along any path, as long as they are kept diametrically opposed at all times. During this trip, A's temperature will change and fluctuate, going up and down, and eventually reach B's temperature. Similarly, B's temperature will also fluctuate, eventually reaching A's temperature. Since the two temperature curves are continuous, there must be a point in time where the two curves meet. When this is reached, the points will be at the same temperature and can be stopped.

A similar a argument shows that it is always possible to find paths where at some point in time, both A and B will be not just at the same temperature but also under the same pressure. If this is not amazing enough, consider that this claim is true not just for temperature and pressure but for any other pair of quantities that vary continuously.

10.11 Short Topics

About bounding boxes. Figure 10.22a shows an image with a minimum bounding box around it. The box is a rectangle. Part (b) of the figure shows the bounding box rotated. It is now bigger, and it is also a square. Part (c) of the figure illustrates why any image can have a square bounding box. As the bounding box of part (c) is rotated, its short sides (in blue) get longer and its long sides (in red) get shorter. The difference between the red and blue sides starts positive and continuously shrinks and ends up being negative. Thus, at a certain point this difference must become zero. At that point, the bounding box becomes a square.

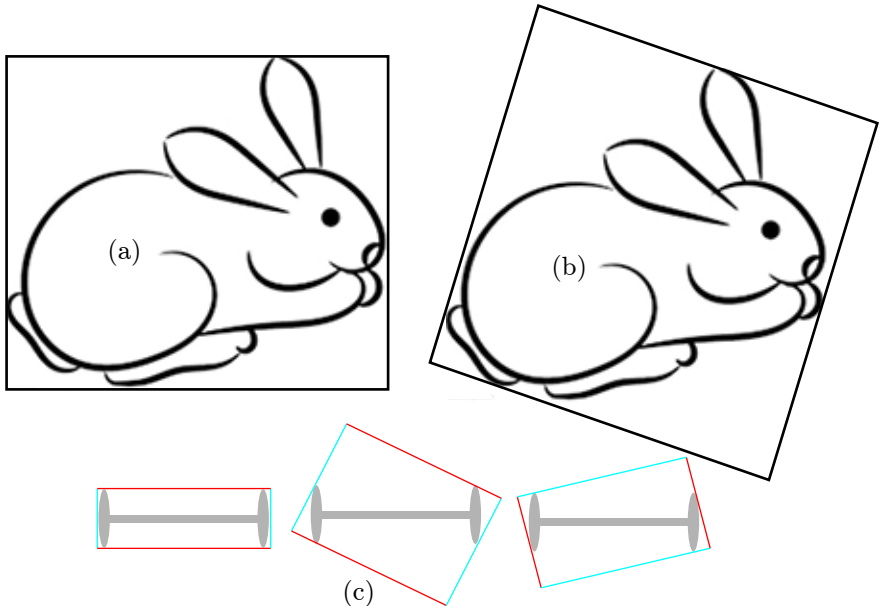


Figure 10.22: Bounding Boxes.

Objects of constant width

A curve of constant width is a closed plane curve whose width, as measured between two bounding parallel lines, is the same in all directions. The simplest example is the circle, whose bounding box is always a square, regardless how much it is rotated. Similarly, a three-dimensional object of constant width is a convex form whose width, measured by the distance between two opposite parallel planes touching its boundary, is the same regardless of the direction of those two parallel planes. (From Wikipedia.)

Figure 10.24 shows two examples. Part (a) of the figure is a Reuleaux triangle. This is a curve of constant width made up of circular arcs of constant radius. It is named after its developer, Franz Reuleaux, and it is constructed from the intersection of three circles (Figure 10.23). General Reuleaux polygons can be similarly formed from any regular polygon with an odd number of sides, or from certain irregular polygons. General Reuleaux polygons can approximate any curve of constant width. Several countries have issued coins shapes like General Reuleaux polygons.

Part (b) of the figure is a regular pentagon where each pair of neighbor vertices is connected with a circular arc, part of a circle whose center is at the vertex opposite the pair. This can be done with any regular polygon with an odd number of sides.

It is also possible to create an arbitrary closed curve of constant width with the following steps (Figure 10.25):

- Draw three intersecting lines. The area bounded by them is an arbitrary triangle. Mark the corner points A, B, and C, as in the figure.
- With point A as the center, draw a circular arc of any radius between segments AB and AC (shown in red).
- Select another point, say C, and draw an arc (in blue) of the same radius between segments CB and CA.
- In a similar way, draw a third arc (in green) centered at B.
- Finally, draw three more arcs of the same radius to complete the closed curve.

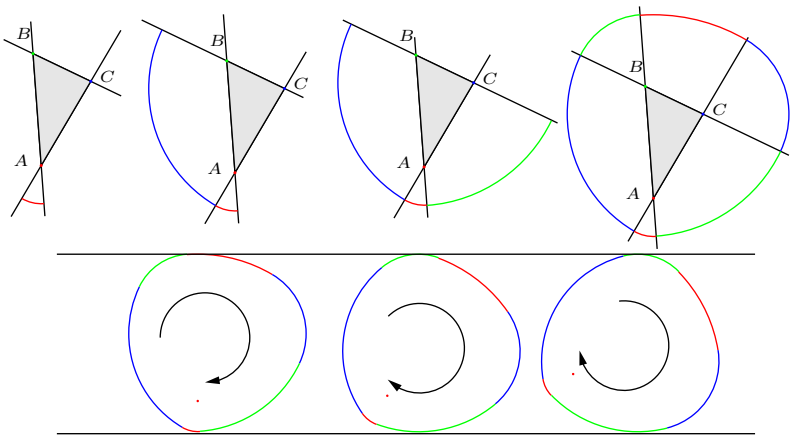


Figure 10.25: An Irregular Closed Curve of Constant Width.

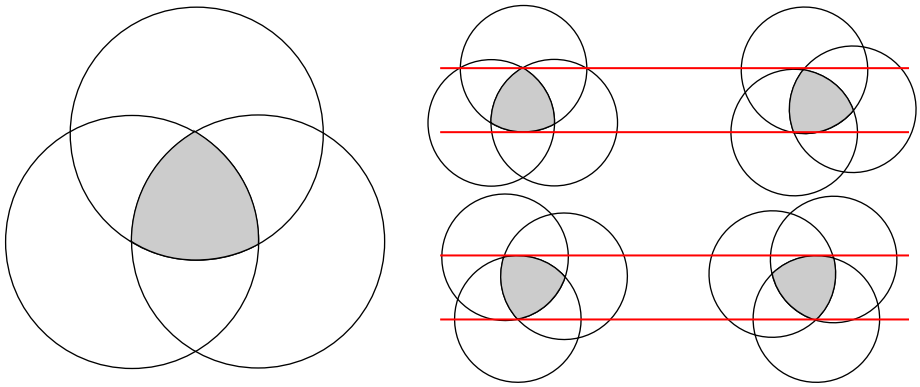


Figure 10.23: Constructing a Reuleaux.

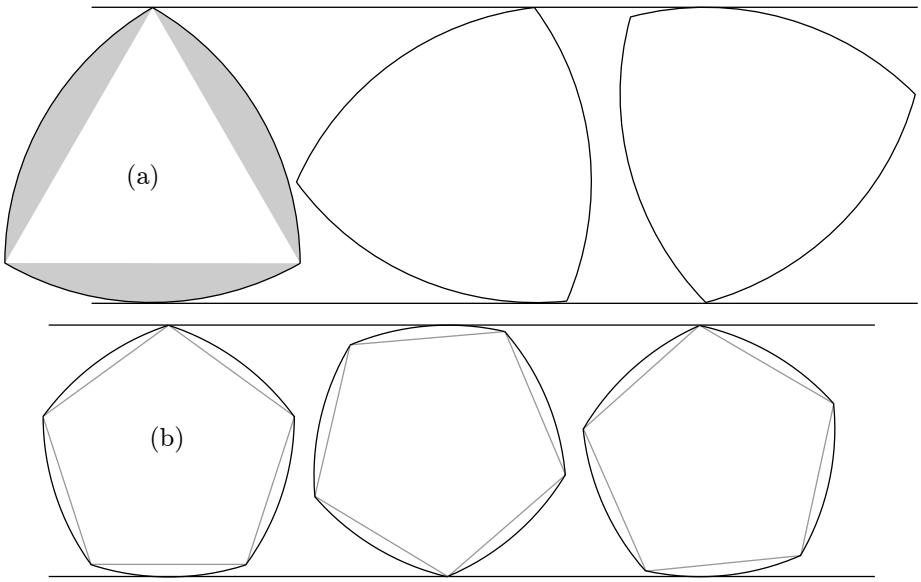


Figure 10.24: Objects of Constant Width.

The width or diameter R of a curve of constant width is defined as the distance between the parallel lines containing the curve. It can be shown that the length of the perimeter of the curve is always πR, the length of the circumference of a circle of diameter R.

This elegant solution of a geometrical problem is discussed, with extensions to the ellipse, in [Blum-Smith 21].

Given a unit circle, we place n equally-spaced points ζ_i on its circumference. Notice that in the Argand diagram, these points become the nth roots of unity. We choose one point and draw chords to connect it with the remaining $n-1$ points. The problem is to compute the product of the lengths of the chords. Figure 10.26a shows the case $n = 7$,

where the six chords radiate from point $\zeta_0 = (1, 0)$. To simplify the problem, we start by connecting each of the n points to the center z of the circle, Figure 10.26b. In this case, each chord is a radius, and it is obvious that its length is given by $|z - \zeta_i|$, the absolute value of the difference of its endpoints. The product of (only six) lengths in this case is therefore

$$P_6(z) = \Pi_{i=1}^{6}|z - \zeta_i| = |\Pi_{i=1}^{6}(z - \zeta_i)|,$$

(we are not including the chord between z and ζ_0.)

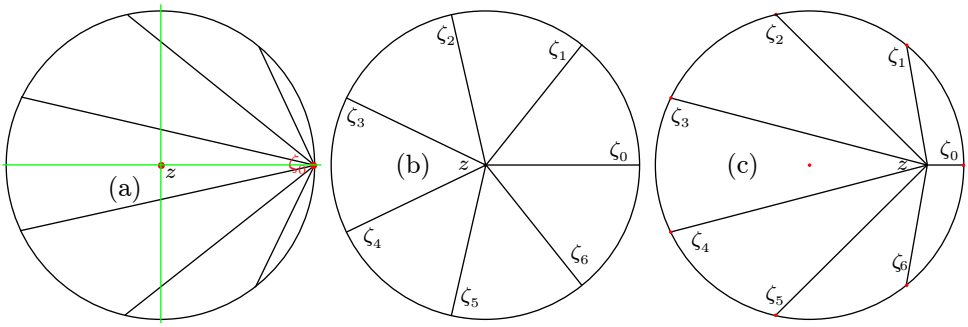


Figure 10.26: Chords in a Circle.

However, we are interested in the product of the lengths of the six chords radiating from ζ_0, and this can now be obtained by moving point z from the center toward ζ_0, dragging the radii with it as shown in Figure 10.26c. In the limit, the seven radii of Figure 10.26b become the six chords whose product we seek. The computation is simplified because the ζ_i are roots of unity, allowing us to write

$$P_6(z) = |\Pi_{i=1}^{6}(z - \zeta_i)|.$$

This becomes especially simple when we multiply $P_6(z)$ by $(z-\zeta_0)$ to obtain $P_7(z) = P_6(z) \times (z - \zeta_0)$, and then divide

$$P_6(z) = \frac{P_7(z)}{(z - \zeta_0)} = \frac{\Pi_{i=1}^{7}(z - \zeta_i)}{(z - \zeta_0)}$$

to keep the same $P_6(z)$. Because of the location of root ζ_0 on the circle, its value is the real number 1, so we write

$$P_6(z) = \frac{P_7(z)}{(z - \zeta_0)} = \frac{\Pi_{i=1}^{7}(z - \zeta_i)}{(z - 1)}.$$

Now that polynomial $P_7(z)$ contains all seven roots of unity, we can write it as $z^7 - 1$. The Mathematica statement `Simplify[Expand[(z^7-1)/(z-1)]]` verifies that $P_6(z) =$

$P_7(z)/(z-1)$ equals $1+z+z^2+z^3+z^4+z^5+z^6$, which is why the absolute value of $P_6(z)$ at $z=1$ is 6, and in general, the product of an $n-1$ chords is n.

The mathematical sciences particularly exhibit order, symmetry, and limitation; and these are the greatest forms of the beautiful.

—Aristotle, *Metaphysica*

Miscellaneous is always the largest category.

—Joel Rosenberg.





Insert: Flags

A national flag is a flag that represents and symbolizes a given nation. It is flown by the government of that nation, but can also be flown by its citizens. A national flag is typically designed with specific meanings for its colors and symbols, which may also be used separately from the flag as a symbol of the nation.

The flag of Nepal is the only national flag which is not a quadrilateral. It also has another unusual, perhaps even unique, property. It is defined rigorously in the Nepal constitution using a geometric construction. (From wikipedia.)

The following are some of the detailed instructions for constructing the flag, copied verbatim from Schedule 1 (Relating to Article 5) of the Nepali constitution:

(A) Method of Making the Shape inside the Border

(1) On the lower portion of a crimson cloth draw a line AB of the required length from left to right.

(2) From A draw a line AC perpendicular to AB making AC equal to AB plus one third AB. From AC mark off D making line AD equal to line AB. Join BD.

(B) Method of Making the Moon

(7) Bisect CF at J and draw a line JK parallel to AB touching CG at point K.

(8) Let L be the point where lines JK and HI cut one another.

(9) Join JG.

(14) With centre M and radius MQ draw a semi-circle on the lower portion touching P and Q.

(17) With centre T and radius TM draw an arc on the upper portion of PNQ touching at two points.

(C) Method of making the Sun

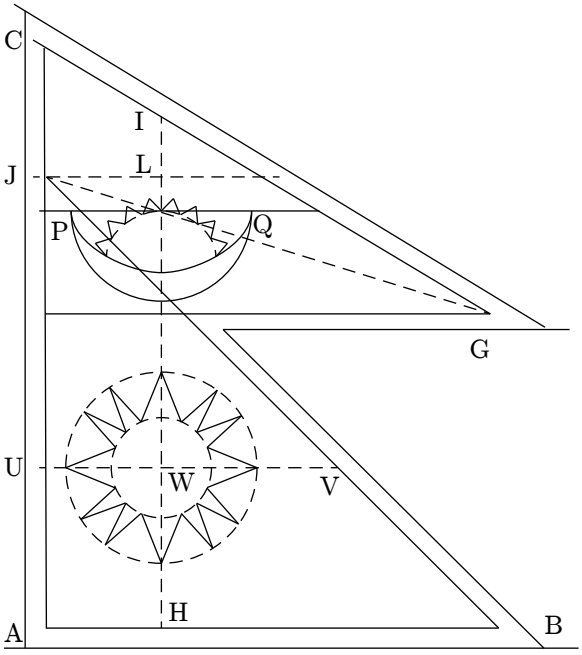
(19) Bisect line AF at U and draw a line UV parallel to line AB touching line BE at V.

(20) With centre W, the point where HI and UV cut one another and radius MN draw a circle.

(D) Method of Making the Border

(23) The width of the border will be equal to the width TN. This will be of deep blue colour and will be provided on all the sides of the flag. However, on the five angles of the flag the external angles will be equal to the internal angles.

Explanation: The lines HI, RS, FE, ED, JG, OQ, JK and UV are imaginary. Similarly, the external and internal circles of the sun and the other arcs except the crescent moon are also imaginary. These are not shown on the flag.







◇ **Exercise 10.6:** What is this exercise doing here?

Bibliography

37everywhere (2024) is tinyurl.com/55p495mb

aatishb (2023) is github.com/aatishb/patterncollider#readme

Aerial (2005) is the video at t.ly/jsCdQ

Agnesi (2024) is witchofagnesi.org

Ahmes (2023) is https://t.ly/Whl17

Alexander, Amir (2014), *Infinitesimal: How a Dangerous Mathematical Theory Shaped the Modern World*, Scientific American / Farrar, Straus and Giroux.

aliquot (2024) is aliquot.de/aliquote.htm

almost (2020) is tinyurl.com/4ucehtce

ambigram (2023) is en.wikipedia.org/wiki/Ambigram

any.number (2023) is tinyurl.com/4p968fpz

Apostol, Tom (2000) "Irrationality of the Square Root of Two—A Geometric Proof," *American Mathematical Monthly*, **107**:841–842, Nov.

arachnoid (2008) is arachnoid.com/is_math_a_science/index.html

Arianrhod, Robyn (2024) *Vector: A Surprising Story of Space, Time, and Mathematical Transformation*, The University of Chicago Press.

Arndt, Jörg and Haenel, Christoph (2006). *Pi Unleashed*, Springer-Verlag. English translation by Catriona and David Lischka.

arxiv (1302) is https://arxiv.org/abs/1302.1479

arxiv (1502) is https://arxiv.org/abs/1502.03501

Bar-Hillel, Maya, and Falk, Ruma (1982), "Some teasers concerning conditional probabilities," *Cognition*, **11** (2): 109–122.

Barnsley, Michael F. (2013) "Fractals Everywhere," *Dover*

Barrow (2011) is `t.ly/KLYFm`

Base 13 function (2024) is `tinyurl.com/3v3cyfsm`

H. Behnke et al. (Eds.) (1974) *Fundamentals of Mathematics*, Volume II, Geometry, The MIT Press Cambridge, Massachusetts.

Benford, Frank (1938) "The Law of Anomalous Numbers," *Proceedings of the American Philosophical Society*, **78**#4, 551–572, March.

Benford (2024) is `en.wikipedia.org/wiki/Benford's_law`

Bergerson (1973) *Palindromes and Anagrams*, Howard W. Bergerson, Dover.

Bergfeldt, Axel (2018) "Open maps which are not continuous," Stack Exchange Mathematics. In an answer to a question. See `tinyurl.com/5bmvtxxn`

Berggren, Lennart, Borwein, Jonathan, and Borwein, Peter (1997), *Pi: a Source Book*, Springer-Verlag.

Berlekamp, Elwyn R. (1968) *Algebraic Coding Theory*, New York, McGraw-Hill.

Bertrand paradox (2024) is `tinyurl.com/4zn68y9a`

Bézier, Pierre (1986) *The Mathematical Basis of the UNISURF CAD System*, Newton, Mass., Butterworth-Heinemann.

Bezier (2024) is `en.wikipedia.org/wiki/Pierre_B%C3%A9zier`

Bezier methods (2024) IS `en.wikipedia.org/wiki/B%C3%A9zier_curve`

Bifurcation (2020) is `tinyurl.com/3r6ymv2s`

BlackSwan (2024) is `tinyurl.com/bdfw69a7`

Blum-Smith, Ben, and Japheth Wood (2021 "Chords of an Ellipse, Lucas Polynomials, and Cubic Equations," *American Mathematical Monthly*, **121**:1, October.

Bogomolny, Alexander (2020) *Cut the Knot*, Amazon.com.

Brain Station Advanced (2025) is `t.ly/gzDpw`

Browne, M. W. (1998), *The New York Times*, Aug. 4.

Calkin, Neil and Herbert S. Wilf (2000) "Recounting the Rationals," *The American Mathematical Monthly*, **107**, #4, 360—363, April.
Available from `http://www.jstor.org/stable/2589182`

Calude, Cristian and Tudor Zamfirescu (1998) "The Typical Number is a Lexicon," *New Zealand Journal of Mathematics*, **27**:7–13.

Campbell and Garnett (1882), Lewis Campbell and William Garnett, *The Life of James Clerk Maxwell*.

Campbell, Joseph (2008) is `http://en.wikipedia.org/wiki/Joseph_Campbell`.

cantor.paradise (2020) is `tinyurl.com/7srx3xp7`

Carus, Paul (1908) "reflections on magic squares," in William Symes Andrews *Magic Squares and Cubes*, Chicago, Open Court, available at `http://www.archive.org/details/magicsquarescube00andrrich`.

Casimir, Hendrik [1983] *Haphazard Reality: Half a Century of Science*, New York, Harper and Row.

ChaosGameNotebk (2024) is `mathworld.wolfram.com/ChaosGame.html`

Cohn, Henry (2006) "A Short Proof of the Simple Continued Fraction Expansion of e," *The American Mathematical Monthly*, **113**, #1, 57–62 (January). Also available from `http://www.jstor.org/stable/27641837`

Conway, John H., et al. (1991) "Geometry and the Imagination," ver. 2.0. Handouts for a two-week summer workshop led by John Conway, Peter Doyle, Jane Gilman, and Bill Thurston at the Geometry Center in Minneapolis, June 17–28. Available from ArXiv.

Conway, John H. and Richard Guy (1996), *The Book of Numbers: Corrected Edition*, Copernicus.

Conway, John H., Heidi Burgiel, and Chaim Goodman-Strauss (2008) "The Symmetries of Things," A K Peters, CRC Press

Conway (2013) is `tinyurl.com/bdf7wezm`

Copeland, A. H. and Paul Erdős, (1946), "Note on normal numbers," *Bulletin of the American Mathematical Society*, **52** (10): 857–860.

Coutteau, Sam (2017) is `tinyurl.com/42a3usxm`

Cummings, Jay (2019) *Real Analysis: A Long-Form Mathematics Textbook*.

datagenetics (2024) is `datagenetics.com/blog/september32012/index.html`

divided.diff (2024) is `en.wikipedia.org/wiki/Divided_differences`

DrBarker (2024) is `youtube.com/watch?v=hDDYYiD6W1A`

du Sautoy, Marcus (2009) "Symmetry, Journey into the Patterns of Nature," Harper Collins.

Duijvestijn, A. J. W. (1978) "A Simple Perfect Square of Lowest Order," *Journal of Combinatorial Theory*, **B-25**(2):240–243.

Dunham, William, (1990) *Journey through Genius: The Great Theorems of Mathematics*, New York, John Wiley.

Edgerton, Samuel (1975), *The Renaissance Rediscovery of Linear Perspective*, NY, Basic Books.

eightfold (2000) Gell-Mann, Murray, and Yuval Ne'eman (2000) *The Eightfold Way*, Westview Press (Paperback).

Einstein tile (2023) is `en.wikipedia.org/wiki/Einstein_problem`

Falbo (2005) is at `jstor.org/stable/30044835`. Published in *The College Mathematics Journal*, **36**#2, 123–134, March.

Feld, Scott L. (1991), "Why your friends have more friends than you do," *American Journal of Sociology*, **96** (6): 1464–1477, available online from `https://pdodds.w3.uvm.edu/research/papers/others/1991/feld1991a.pdf`

Feller, William (1968), "An Introduction to Probability Theory and Its Applications," John Wiley, 3rd Edition, Vol. I, p. 251.

Fewster, R, M. (2009) "A Simple Explanation of Benford's Law," *The American Statistician*, **63**(1):26–32, February.

FibonacciMusic (2022) is `tinyurl.com/bdzks4zn`

Flajolet, Philippe, and Martin, G. Nigel (1985). "Probabilistic counting algorithms for data base applications," *Journal of Computer and System Sciences 31*, **31**(2), pp. 182–209, October.

Flanders, Harley (1999) "Math Bite: Irrationality of $\sqrt{m}$," *Mathematics Magazine*, **72**(3) p. 235, June.

Fowler, D. H. (1981) "A Generalization of the Golden Section," *Fibonacci Quarterly*, **20**, 146–158.

Fraenkel, Aviezri and Shmuel Klein (1996). "Robust universal complete codes for transmission and compression." *Discrete Applied Mathematics*, **64** (1): 31–55.

funny.sequences (2018) is `tinyurl.com/mwbj7jcx`

G.root5 (2018) is `tinyurl.com/2tscn6mp` by Matt Parker.

gamma.history (2024) is `tinyurl.com/2d5n7ay2`

Gamow, George (1961) *One, Two, three,... Infinity*, New York, Viking, 1961.

GoldenQuartics (2020) is `xperimex.com/blog/golden-quartics`

Gould (2020) is `tinyurl.com/mw6zb9ep`

Grand-Hotel (2001) is the wikipedia article `Hilbert's_paradox_of_the_Grand_Hotel`.

GregSmith (2015) is `tinyurl.com/4payt9wy`

Grime (2015) is `https://t.ly/oI4UG`

Guilbeau , Lucye (1930) "The History of the Solution of the Cubic Equation," *Mathematics News Letter* **5**(4):8–12.

Guy (1988). "The strong law of small numbers," *Am. Math. Mon.* **95** (8): 697–712.

Guy (1990) "The Second Strong Law of Small Numbers," *Math. Mag.*, **63**, 3–20.

Hanson, Andrew J. (2006) *Visualizing Quaternions*, San Francisco, Elsevier Morgan-Kaufmann.

Hill, Theodore (1995), "A Statistical Derivation of the Significant-Digit Law," *Statist. Sci.*, **10** #4, 354–363, November. Available from `tinyurl.com/bdefvaea`

Hill, Theodore P. (1995) "Base-invariance implies Benford's law," *Proceedings of the A. M. S.*, **123**, 887-895.

hotbits (2024) is `fourmilab.ch/hotbits/`

Hue (2013) is `tinyurl.com/26fs4yzm`

Huntley, H. E. (1970) *The Divine Proportion: A Study in Mathematical Beauty*, New York, Dover Publications.

Hutchinson, John E. (1981). "Fractals and self similarity," *Indiana Univ. Mathematics Journal*, **30** #5, 713–747. Available from `tinyurl.com/29aju92t`

ImpossibleSquares (2020) is `tinyurl.com/3yw4uzys`

IntegerProport (2023) is `tinyurl.com/yck232k9`

iOrnament (2024) is `science-to-touch.com`

Jarvis, P. (1990) "Implementing CORDIC Algorithms," *Dr. Dobb's Journal*, 152–158, October,

Kac, Mark (1984), "Marginalia: more on randomness," *American Scientist*, **72** (3): 282–283, May–June. Available online at `https://tinyurl.com/mwejs6xk`

Kac, Mark (1985) *Enigmas of Chance: An Autobiography*, Harper & Row.

Keller, Joseph B. (1986) "The Probability of Heads," *The American Mathematical Monthly*, **93**#3, 191–197, March.

Keller, Joseph B. et al., (1989) "Fair Dice," *American Mathematical Monthly*, **96**, 337–339.

kexue.fm (2023) is `tinyurl.com/yahtn7z5`

Kimball (2006) "Teachings of Presidents of the Church: Spencer W. Kimball," The Church of Jesus Christ of Latter-day Saints.

King, Ross (2000) "Brunelleschi's Dome: How a Renaissance Genius Reinvented Architecture," New York, Walker and Company; London, Chatto and Windus.

Knuth, Donald E. (1969) *The Art of Computer Programming*, Vol. 2, "Seminumerical Algorithms," Addison-Wesley, Reading MA, p. 220.

Knuth, Donald E. (1976) "Coping With Finiteness," *Science*, **194**(4271):1235–1242, December.

Gregg Kreutz (2001) *Camera Absurda: The Case against Hockney*, Penguin Putnam inc.

Kucharski, Adam (2014) "Math's Beautiful Monsters," available online at `tinyurl.com/bde82684`

L-systems (2024) is `en.wikipedia.org/wiki/L-system`

Lanier (2009) is `jwilson.coe.uga.edu/emt725/Class/Lanier/Nine.Digit/nine.html`.

Lines That Connect (2021) is `tinyurl.com/bdz78vm4`

Livio, Mario (2005) *The Equation That Couldn't Be Solved: How Mathematical Genius Discovered the Language of Symmetry*, Simon and Schuster.

Livio, Mario (2024), The Golden Ratio in Art: Drawing heavily from The Golden Ratio. `www.math.ksu.edu/~cjbalm/Quest/Breakouts/GR.pdf`

logistic map (2024) is `en.wikipedia.org/wiki/Logistic_map`

Loomis, Elisha (1968) *The Pythagorean Proposition: Its Demonstration Analyzed and Classified*, National Council of Teachers of Mathematics, Washington, DC. This is also freely available on the Internet in DjVu format.

Mandalian, Tigran (2024) *Fibonacci Numbers Shape the Duration of Historical Progress: Predicting the Future*, published by Tigran Mandalian.

Manetti, Antonio (1488) *The Life of Brunelleschi*, translated by Catherine Enggass, University Park, PA, Pennsylvania State University Press, 1970.

Maor, Eli (1987) *To Infinity and Beyond: A Cultural History of the Infinite*, Princeton University Press.

Maor, Eli (2007) *The Pythagorean Theorem: A 4,000-Year History*, Princeton, NJ, Princeton University Press.

Mathellaneous (2008) is `math.stanford.edu/~mathews /articles/mathellaneous1.pdf`.

mathforum (2011) is `tinyurl.com/9rxc88pz`

Mathoma (2015) is `https://t.ly/6UvHX`

Mathoma (2016) is `https://t.ly/EQF2W`

MathPages (2009) is `mathpages.com/home/kmath028/kmath028.htm`.

MathsNeedNotApply (2025) is `https://t.ly/9gMBk`

Mathworld (2005) is `http://mathworld.wolfram.com/Quaternion.html`.

math.wrath (2023) is `tinyurl.com/3uwjsd2c`

May, C. P. (1962) *James Clerk Maxwell and Electromagnetism*, New York, Franklin Watts.

M. D. Meyerson (1996) "The x^x Spindle," *Mathematics Magazine*, **69**(3), pp. 198–206.

MindYourDecisions (2017) is `t.ly/Gypi5`

MindYourDecisions (2023) is `t.ly/9YrqL`

MindYourDecisions (2024) is `t.ly/NgvCD`

MindYourDecisions (2025) is `https://t.ly/01Yvh`

minutephysics (2022) is `youtube.com/watch?v=-eqdj63nEr4`

Moire (2015) is `t.ly/P-6gU`

Moire (2025) is `sqrt.ch/museum`

Morphocular (2021) is `t.ly/VEHvG`

Morphocular (2025) is `https://t.ly/RjtT6`

mostIrrat (2023) is `tinyurl.com/46tcbtc5`

Mr H Tutoring (2025) is `https://t.ly/a7lI3`

Mufano (2009) is `mrob.com/pub/math/largenum.html`.

Nadis (2023) is `tinyurl.com/mwvucexk`

Newcomb, Simon (1881) "Note on the Frequency of Use of the Different Digits in Natural Numbers," *American Journal of Mathematics*, **4**, 39–40.

nonlinearLS (2024) is `rb.gy/pakrwc`

number196 (2015) is `youtube.com/watch?v=bN8PE3eljdA`

numberphile (2017) is `t.ly/OPG_W`

Numeropedia (2009) is `numeropedia3.googlepages.com/numbers10t`.

OptWhiz (2023) is `tinyurl.com/2s43cz36`

Peek, J.B.H. (2009) *Origins and Successors of the Compact Disc*, Springer Verlag.

Penelope (2009) is `samuel-beckett.net/Penelope/four_symbolism.html`.

PenguinMaths (2020) is `https://t.ly/F7exl`

Perlin, K. (1985), "An image synthesizer," *ACM SIGGRAPH '85*, **19** (3): 287–296.

Perlin (1999) "Making Noise," `noisemachine.com/talk1`

Perlin, K., and F. Neyret (2001) "Flow noise," *SIGGRAPH Technical Sketches and Applications*.

Perlin 1D (2024) is `t.ly/VOpfI`

Pi5 (2020) is `tinyurl.com/yeyn7899`

PiCollision (2019) is `tinyurl.com/mr2cf2pe`

Pohlmann, Ken C. (1992) *The Compact Disc Handbook*, 2nd edition, A-R Editions, Inc.

pomax (2020) is `pomax.github.io/bezierinfo`

Posamentier, Alfred S., and Lehmann, Ingmar (2004), *π: A Biography of the World's Most Mysterious Number*, Prometheus Books.

Poundstone, William (1995) *Prisoner's Dilemma*, Anchor books.

Quine (2023) is `tinyurl.com/bd9brj3m`

Rabinowitz, Stanley and Stan Wagon (1995), "A Spigot Algorithm for the Digits of π," *The American Mathematical Monthly*, **102**(3), 195–203, March. Available to subscribers at `tinyurl.com/mvxpkj37`

ramsey.bazett (2023) is `tinyurl.com/33wdmtf6`

redditI (2020) is `https://tinyurl.com/8paexf8r`

Reed, Irving S., and Gustave Solomon (1960) "Polynomial Codes over Certain Finite Fields," *SIAM Journal of Applied Mathematics*, **8**(10):300–304.

Reid (2004). David Reid, "Teaching Mathematics through Brick Patterns," *Nexus Network Journal*, **6**(2).

Renteln, Paul and Allan Dundes (2005) "Foolproof: A Sampling of Mathematical Folk Humor," *Notices of the AMS*, **52**(1):24–34.
Available at `http://www.ams.org/notices/200501/fea-dundes.pdf`.

Rhombictricontahedron (2024) is `tinyurl.com/69mjrvz7`

Saalman, Howard (1970), *The Life of Brunelleschi, by Antonio Tuccio Manetti*, edited. Penn State University.

Saari, Donald G. and Zhihong (Jeff) Xia (1995) "Off to Infinity in Finite Time," *Notices of the AMS*, **42**, #5, 538–546, April

Salomon, David (2011) "The Computer Graphics Manual," London, Springer Verlag.

Schroeder, M. (1991) "Noble and Near Noble Numbers." In *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. New York: W. H. Freeman, pp. 392–394.

seven magazine (2009) is `http://www.sevenmagazine.org/`.

Sgrilli, Bernardo Sansone (1733) *Descrizione e studi dell'insigne fabbrica di Santa Maria del Fiore metropolitana fiorentina*, Florence, Bernardo Paperini.

Shannon (2025) is `t.ly/gp8aq`

Sherman, B. F. (1993), "The fourth side of a triangle," *Math. Mag.*, **66**, 333–337.

sine.amazing (2024) is `youtube.com/watch?v=vlXELKCtJ-4`

Sloane, Neal (2008) *The On-Line Encyclopedia of Integer Sequences*, located at `eois.org`.

Smith (2023) is `hedraweb.wordpress.com`

spacegoo (2023) is `http://www.spacegoo.com/penrose`

stackexchange1851698 (2023) is `tinyurl.com/ytyyaebp`

stackexchange4856811 (2024) is `tinyurl.com/2p8j5w56`

Stølum, Hans-Henrik (1996) "River Meandering as a Self-Organization Process," *Science*, **271**, issue 5256, 1710–1713, March.

Su, Francis E., et al. (2025) "Sierpinski-Mazurkiewicz Paradox." *Math Fun Facts*, `https://www.math.hmc.edu/funfacts`.

SuperFactorial (2021) is `tinyurl.com/bdsz3nyv`

Swartzlander, Earl E. (1990) *Computer Arithmetic*, Silver Spring, MD, IEEE Computer Society Press.

Symonds (2014) is `tinyurl.com/yw29tr8z`

Taubin, Gabriel (1994) "An accurate algorithm for rasterizing algebraic curves and surfaces," *IEEE Computer Graphics and Applications*, **14**(March):14–23.

Thurston, William (1980) *The Geometry and Topology of Three-Manifolds*, Princeton University lecture notes.

tomverbeure (2022) is `tinyurl.com/372fbz6f`

Trahan, Alexander C., (2005) "When Does Randomness End?" is `t.ly/dpxI1`

Tupper, Jeff (2001), "Reliable Two-Dimensional Graphing Methods for Mathematical Formulae with Two Free Variables," *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, August 2001, Pages 77–86. `https://doi.org/10.1145/383259.383267`

Tyson, Neil deGrasse (2007) *Death by Black Hole*, N.Y., W. W. Norton.

Vachss, Raymond (1987) "The CORDIC Magnification Function," *IEEE Micro*, **7**(5)83–84, October.

veritasium (2025) is `t.ly/8gogd`

Vitali (2025) is `t.ly/wGc5D`

vivint (2017) is `tinyurl.com/2tvkmzev`

Vogel, Helmut (1979), "A better way to construct the sunflower head," *Mathematical Biosciences*, **44**: 179–189.

Volder, Jack E. (1959) "The CORDIC Trigonometric Computing Technique," *IRE Transactions on Electronic Computers*, **EC-8**:330–334.

Von Neumann, John (1956) *Probabilistic Logics and the Synthesis of Reliable Organism from Unreliable Components*, Automata Studies, Annals of Mathematics, Princeton, pp. 43–98.

Walker, Paul Robert (2002) *The Feud That Sparked the Renaissance: How Brunelleschi and Ghiberti Changed the Art World*, New York, HarperCollins.

Walther, John S. (1971) "A Unified Algorithm for Elementary Functions," *Proceedings of Spring Joint Computer Conference*, **38**:379–385.

WeirdFairDice (2016) is `tinyurl.com/me3w8wf5`

Weisstein, Eric W. (2024) "Weierstrass Function." From MathWorld–A Wolfram Web Resource. `https://mathworld.wolfram.com/WeierstrassFunction.html`

Weisstein-pickin (2007) is Weisstein, Eric W. "Real Number Picking." From MathWorld, A Wolfram web resource. `mathworld.wolfram.com/RealNumberPicking.html`.

Weisstein-Ziv (2023) is Weisstein, Eric W."Pi Approximations." From *MathWorld–A Wolfram Web Resource*.
`mathworld.wolfram.com/PiApproximations.html`

Wells, David (1998) *The Penguin Book of Curious and Interesting Numbers*, New York, Penguin publishing.

Wigner, Eugene (1960) "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," *Communications on Pure and Applied Mathematics* **13**(1):1–14.

WikiAristoPardox (2023) is `en.wikipedia.org/wiki/Aristotle's_wheel_paradox`

WikiBBP is `tinyurl.com/52vwd7bu`

WikiBifurcat (2023) is `en.wikipedia.org/wiki/Bifurcation_diagram`

Wikibirthday (2023) is `https://en.wikipedia.org/w/index.php?title=Birthday_problem&oldid=481710309`

WikiComplex (2023) is `tinyurl.com/mr27c6c8`

WikiDensty (2023) is `tinyurl.com/48ubm4cs`

WikiDirichlet (2024) is `en.wikipedia.org/wiki/Dirichlet_function`

WikiFibon (2023) is `tinyurl.com/yrz39fcb`

WikiHash (2024) is `en.wikipedia.org/wiki/Hash_table`

WikiHorn (2023) is `en.wikipedia.org/wiki/Gabriel's_horn`

WikiInteresting (2009) is `en.wikipedia.org/wiki/Interesting_number_paradox`

WikiMonty (2023) is `https://en.wikipedia.org/wiki/Monty_Hall_problem`

WikiOmega (2024) is `en.wikipedia.org/wiki/Chaitin%27s_constant`

WikiParadoxes (2024) is `en.wikipedia.org/wiki/List_of_paradoxes`

WikiPhi (2023) is `tinyurl.com/yaahrhdr`

WikiPotato (2025) is `en.wikipedia.org/wiki/Potato_paradox`

WikiQuadr (2024) is`en.wikipedia.org/wiki/Quadratic_formula`

WikiQuaternion (2025) is `http://en.wikipedia.org/wiki/Quaternion`.

WikiRodrigues (2025) is `https://t.ly/C1e93`

WikiSpline (2024) is `en.wikipedia.org/wiki/Spline_(mathematics)`

WikiTetrat (2023) is `tinyurl.com/2tkxp9ky`

WikiWigner (2023) is `tinyurl.com/48zbch4c`

wolfram.platonic (2024) is `mathworld.wolfram.com/PlatonicSolid.html`

Xia, Zhihong (1992) "The Existence of Non-collision Singularities in Newtonian Systems," *Annals of Mathematics*, **135**, #3, 411–468, May.

XOR (2023) is `en.wikipedia.org/wiki/Exclusive_or`

Yiu, Paul (2012) "Sherman's Fourth Side of a Triangle," *Forum Geometricorum* **12**, 219–225. Available from `tinyurl.com/yfnnvmfj`

youtube (2023) is `tinyurl.com/ypruf6t8`

youtube.gen.fibo (2024) is `tinyurl.com/4ekz8jha`

YoutubePeyam (2020) is `tinyurl.com/3wszy3ch`

youtube.pisano (2013) is `tinyurl.com/mr2zxdnp`

youtube.twit (2024) is `youtube.com/watch?v=Qj85v2ca89E`

yt.ant.string (2023) is `tinyurl.com/y5555y7n`

But it's not just a game of finding literary references.
—Dan Simmons *(1938)*.

Answers to Exercises

> A bird does not sing because he has an answer, he sings because he has a song.
>
> —Chinese Proverb.

1. (Tongue in cheek.) Eliminating one equation would have doubled the sales from one (2^0) to two (2^1) copies. Eliminating one more equation would have sold four (2^2) copies. Thus, eliminating all 20 equations would have resulted in $2^{20} = 1{,}048{,}576$ copies sold. Among other things, this also explains the recent revival of interest in the writings of Jane Austen.

2. The proof is by contradiction. We assume that the cube root of 2 is rational and equals the ratio a/b of two integers. The steps are:

$$\sqrt[3]{2} = \frac{a}{b} \rightarrow 2 = \frac{a^3}{b^3} \rightarrow b^3 + b^3 = a^3,$$

and the result contradicts Fermat's Last Theorem, since a and b are integers.

1.1. This surface consists of a set of circles of radius p in the yz plane. The x coordinate is simply w, so each circle is drawn slightly to the right of its predecessor. The radius p of a circle varies periodically with the x coordinate w, leading to larger circles, followed by smaller ones, followed again by larger ones, and so on (Figure Ans.1).

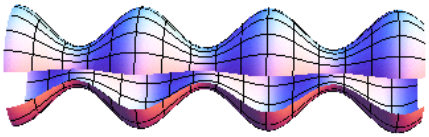
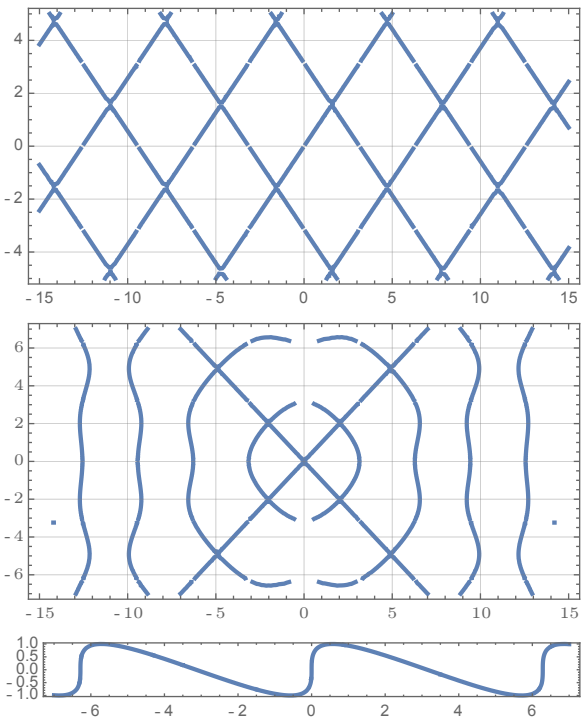


Figure Ans.1: Varying The Radius.




```
ContourPlot[ y Sin[x]/Sin[y]==y, {x,-15.,15}, {y,-5,5},Frame->True,
 GridLines->Automatic, AspectRatio->.5]
ContourPlot[x Sin[x]/Sin[y]==y, {x,-15.,15}, {y,-7,7},Frame->True,
 GridLines->Automatic, AspectRatio->.5]
ContourPlot[Sin[x+y]==y, {x,-7,7}, {y,-1,1}, AspectRatio->.1]
```

Figure Ans.2: Three Implicit Sine Functions.

1.2. Figure Ans.2 plots three such implicit functions and lists their Mathematica codes.

1.3. The basic pattern is a white square whose top-left half is a colored triangle and has a small colored circle at the bottom-right corner.

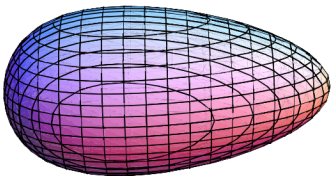
1.4. Make a pool table in the form of an ellipse with a hole at one focus. Place a billiard ball at the other focus and hit it toward the edge of the table in any direction. The ball would hit the edge and be reflected toward the other focus.

1.5. A two-dimensional oval can be obtained from an ellipse by perturbing one coordinate, which suggests that a three-dimensional egg shape can be obtained from an ellipsoid by perturbing two coordinates in the same way. The equation of an ellipsoid is $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$, which is why the shape of an egg can be obtained by an expression of the form $g(x/a)^2 + g(y/b)^2 + (z/c)^2 = 1$, where g can be $1 + 0.25z$ or a similar small perturbation. The code

```
a=1; b=1; c=2; g=1+0.25z;
```

```
ContourPlot3D[g(x/a)^2+g(y/b)^2+(z/c)^2==1, {x,-a-0.2,a+0.2},
 {y,-b-0.2, b+0.2}, {z,-c,c}, Axes->False, BoxRatios->Automatic]
```

produces the elegant egg



2.1. The idea that logical operations (or Boolean algebra) can serve as the basis of implementing arithmetic operations was the brainchild, in 1935–36, of Claude Shannon. In his 1936 master's thesis, reference [Shannon 25] published in 1938, he applied logical (Boolean) operations to the design of switching circuits, which inspired later computing pioneers to use binary numbers in their computer designs. Today, this work is believed by many to be the most important master's thesis of the twentieth century.

The first electronic computers appeared in the 1940's, during and after World War II. Most were special-purpose machines, designed for code breaking, whose developers did not realize the advantages of binary numbers. Today, we believe that the first binary computer was the Zuse Z1, a motor-driven mechanical computer designed by Konrad Zuse from 1936 to 1937. The first electronic binary computer (although not general-purpose) was the 1941 Atanasoff-Berry Computer (ABC), built by John Vincent Atanasoff and Clifford E. Berry.

2.2. Because the ancient Romans did not have the concept of zero. They started the first century at year 1, so it had to end at year 100, implying that the 2nd century started at year 101, and so on.

2.3. No, because any odd number has this property.

2.4. Let's look at the simple case of a 3-digit integer. Given the integer $100a+10b+c$, if the sum of its digits is divisible by 3, we can write $a+b+c=3x$. Substituting yields

$$100(3x-b-c)+10b+c=300x-90b-99c.$$

This is a three-part sum where each part is divisible by 3. This method can be applied to n-digit numbers. (See also Exercise 2.7.)

2.5. The sum of the digits of such a number is $1+6+1+6+1=15$, so it is divisible by 3 and therefore composite.

2.6. Natural hexagons are found in: Snowflakes, certain ice crystals, the compound eyes of various insects, benzene rings of six carbon atoms and other cyclic compounds, basalt rocks, quartz, marine skeletons, the edges of soap bubbles that are packed together, an enormous spinning hexagon was discovered in the 1980's in the clouds at Saturn's north pole, the country of France is often referred to in French as l'Hexagone, plant cells of the

moss Plagiomnium affine, a spider's web, pink lily and tulip have six leaves, the pillars of Fingal's Cave (Scotland) and of the Giant's Causeway (in UK and Ireland) have a hexagonal cross section.

Artificial hexagons are found in the white divisions of a volleyball, the patterns of a basketball, a stop sign, pencils, the heads of bolts, nuts, floor and wall tiles as well as paving stones, crystals in rhombic shape, the boards of certain games, the folds of a hexaflexagon, the reflections inside a kaleidoscope, the components of the primary mirror of the James Web Space Telescope.

2.7. Let's look at the simple case of a 3-digit integer. Given the integer $100a+10b+c$, if the sum of its digits is divisible by 9, we can write $a+b+c=9x$. Substituting yields

$$100(9x-b-c)+10b+c=900x-90b-99c.$$

This is a three-part sum where each part is divisible by 9. This method can be applied to any n-digit numbers.

Here is an alternative proof. Given an arbitrary $(n+1)$-digit integer $a_n a_{n-1}\ldots a_1 a_0$, we write it explicitly

$$N=10^n a_n+10^{n-1}a_{n-1}+\cdots+10^1 a_1+10^0 a_0,$$

and denote the sum $a_n+a_{n-1}+\cdots+a_1+a_0$ of its digits by $T(N)$. We now subtract

$$N-T(N)=(10^N-1)a_n+(10^{n-1}-a_{n-1})+\cdots+(10-1)a_1 \stackrel{\text{def}}{=} B_n a_n+B_{n-1}a_{n-1}+\cdots+B_1 a_1.$$

We notice that each coefficient B_i is of the form $999\ldots9$ and is therefore divisible by 9. Thus, the right-hand side of the above equation is divisible by 9, which shows that N itself is divisible by 9 if the sum $T(N)$ on the left side has the same property. This also shows that N is divisible by 3 if $T(N)$ is so divisible.

2.8. Because π is defined in connection with circles, whereas this series is the sum of (inverse) squares.

2.9. The next element is 111 (more accurately "one hundred and eleven"). The kth element of this sequence is the smallest nonnegative integer whose name in U.S. English consists of k syllables.

As an aside, the name of the integer 111,777 is "one-hundred and eleven thousand, seven-hundred and seventy seven," and this is the smallest integer whose name consists of 19 syllables $(1+2+1+3+2+2+2+1+3+2=19)$. This can be cut to 18 syllables by omitting the second "and" in the name, or by using the following, 18-syllable phrase to describe this integer "the least integer not nameable in fewer than nineteen syllables" $(1+1+3+1+3+1+2+1+2+3=18)$. See also Berry's Paradox.

2.10. Current Internet resources and sophisticated mathematical software make this easy. The two pairs for 1,729 are $(1,12)$ and $(9,10)$. The two pairs for 635,318,657 (discovered by Euler in 1772) are $(133,134)$ and $(158,59)$.

2.11. A simple search yields the short video at reference [number196 15].

2.12. The number is 10|31|22|33|14.

2.13. The phrase "love of God" yields $12+15+22+5+15+6+7+15+4=101\%$, which tells us that hard work and knowledge should be everyone's goal, attitude is even more important, but love of God is what our peers mean when they want us to give 101%.

2.14. The Copeland-Erdős constant is normal in base 10. This number, described in [Copeland 46], is 0.235711131719232931374143475359616771737983 89 . . .. It is constructed by concatenating the prime numbers in base 10.

2.15. Yes. To construct right-truncatable primes, we start again with the four 1-digit primes 2, 3, 5, and 7. We append more digits to the right of each, check for primality, and continue until one number is left.

Here are the first few such numbers, together with the largest one 2, 3, 5, 7, 23, 29, 31, 37, 53, 59, 71, 73, 79, 233, 239, 293, 313, . . . , 73939133 (sequence A024770 in the OEIS, reference [Sloane 08]).

2.16. It seems that the first three surprising relations are coincidences, but the length 6 sequence generated by compressing the powers of 2 can perhaps be explained.

2.17. This is simple. Assume that $\sqrt{p}$ is rational and it equals the ratio of integers a/b. Then $b\sqrt{p}=a$ and

$$b^2p=a^2. \tag{Ans.1}$$

Recall that the prime factorization theorem, also known as the Fundamental Theorem of Arithmetic, states that every positive integer greater than 1 can be expressed uniquely as a product of prime numbers. Thus, if the prime factorization of x is $p_1p_2p_3$, then the prime factorization of x^2 must be $p_1p_1p_2p_2p_3p_3$; it has an even number of factors. Applying this to Equation (Ans.1), we see that its left-hand side has an odd number of prime factors and its right-hand side has an even number of factors. This contradiction proves the claim.

It is easy to prove in a similar way that if a is not a perfect square, then $\sqrt{a}$ is irrational.

2.18. All the square roots of non-square integers are irrational, as are many logarithms. Perhaps the most well-known irrational, after π, e, and $\sqrt{2}$, is the golden ratio, $\varphi\approx 1.618$. See also Table 2.16.

2.19. No. It is unknown whether π contains every possible pattern, but it is easy to construct irrational numbers that consist of only a certain type of pattern, such as the number

0.0 1 00 11 000 111 0000 1111 00000 11111 . . ..

2.20. Many sequences and series whose terms are rational converge to an irrational limit. Here are some examples:

- The irrational Euler's number e is one of the important mathematical constants. One of its many definitions is the limit of the infinite sequence

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n,$$

of rational terms.

- The infinite sequence of ratios of consecutive Fibonacci numbers F_{n+1}/F_n converges to the irrational golden ratio ϕ.

- Khinchin's constant, Page 62, is the limit of the geometric mean of the (integer) coefficients of the continued fraction representations of almost all real numbers. Its value is 2.6854..., and it is "suspected" to be irrational.

- The Euler series, Equation (5.15), converges to $\pi^2/6 \approx 1.644$.

2.21. We choose the two irrationals $x = y = \sqrt{2}$. If $x^y = \sqrt{2}^{\sqrt{2}}$ is rational, we are done. Otherwise, $\sqrt{2}^{\sqrt{2}}$ must be irrational, which permits us to choose $x = \sqrt{2}^{\sqrt{2}}$ and $y = \sqrt{2}$, but this also ends up with an x^y that is rational, because

$$x^y = \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \left(\sqrt{2}\right)^{\sqrt{2}\cdot\sqrt{2}} = \left(\sqrt{2}\right)^2 = 2.$$

2.22. This basic proof is by contradiction. We first assume that $\tan 1^\circ$ is rational, then prove from this that $\tan 2^\circ$, $\tan 3^\circ$ is, and so on, are all rationals. However, the right triangle with sides 1 and $\sqrt{3}$ has a hypothenuse of $\sqrt{1^2+3} = 2$. The angle between its 1-unit side and the 2-unit hypothenuse is 60° (the other angles are 30° and 90°), which implies that $\tan 60^\circ = \sqrt{3}/1$ and is therefore irrational. A contradiction.

One of the many trigonometric identities is

$$\tan(\alpha+\beta) = \frac{\tan\alpha + \tan\beta}{1-\tan\alpha\tan\beta}.$$

When this is applied to $\alpha = \beta = 1^\circ$, we get

$$\begin{aligned}\tan 2^\circ &= \tan(1^\circ + 1^\circ)\\ &= \frac{\tan 1^\circ + \tan 1^\circ}{1-\tan 1^\circ \tan 1^\circ}\\ &= \frac{\text{rational} + \text{rational}}{1 - \text{rational}}\\ &= \text{rational}.\end{aligned}$$

Hence, assuming that $\tan 1^\circ$ is rational implies that $\tan 2^\circ$ is also rational, which in turn implies that ....

2.23. Assume that $p \stackrel{\text{def}}{=} \sqrt{2}+\pi$ is rational, and consider the quadratic equation $x^2-2px+(p^2-2)=0$. This algebraic equation has rational coefficients, and it's easy to verify that one of its solutions is π. But π has been proved to be transcendental, which is a contradiction. Here is a simple check:

$$\pi^2-2(\sqrt{2}+\pi)\pi+[(\sqrt{2}+\pi)^2-2]=\pi^2-2\pi\sqrt{2}-2\pi^2+2+2\pi\sqrt{2}+\pi^2-2=0.$$

2.24. Start by considering the identity

$$(\sqrt{2}+1)(\sqrt{2}-1)=(\sqrt{2})^2-1^2=1\rightarrow\frac{1}{\sqrt{2}-1}=\sqrt{2}+1=2.414213\ldots,$$

which we use to start the continued fraction

$$\sqrt{2}=1+.414213\ldots=1+\frac{1}{\frac{1}{.414213\ldots}}=1+\frac{1}{2+.414213\ldots}=1+\frac{1}{2+\frac{1}{\frac{1}{.414213\ldots}}},$$

and continue in this way to end up with

$$\sqrt{2}=1+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\ddots}}}}.$$

This method can easily be applied to the continued fraction expansion of other square roots, as the following examples illustrates. We start with $\sqrt{5}$.

$$\begin{aligned}
\sqrt{5}&=2+\frac{1}{x}.\\
\sqrt{5}-2&=\frac{1}{x}.\\
x=\frac{1}{\sqrt{5}-2}&=\frac{1}{\sqrt{5}-2}\frac{\sqrt{5}+2}{\sqrt{5}+2}=\frac{\sqrt{5}+2}{5-4}=\sqrt{5}+2.\\
&=2+\frac{1}{x}=2+\frac{1}{2+\sqrt{5}}=2+\frac{1}{2+2+\frac{1}{x}}=2+\frac{1}{4+\frac{1}{2+\sqrt{5}}}=\ldots.
\end{aligned}$$

which leads to the expansion $\sqrt{5}=[2;4,4,4,\ldots]$.

The case of $\sqrt{3}$ is slightly more complex.

$$\begin{aligned}
\sqrt{3}&=1+\frac{1}{x},\quad \sqrt{3}-1=\frac{1}{x}.\\
x=\frac{1}{\sqrt{3}-1}&=\frac{1}{\sqrt{3}-1}\frac{\sqrt{3}+1}{\sqrt{3}+1}=\frac{\sqrt{3}+1}{3-1}=\frac{\sqrt{3}+1}{2}.
\end{aligned}$$

We now argue that $\sqrt{3}$ is about 1.7, which is why $(\sqrt{3}+1)/2$ is a little greater than 1, and we write it as

$$\frac{\sqrt{3}+1}{2} \stackrel{\text{def}}{=} 1+\frac{1}{y},$$

which permits us to write

$$\sqrt{3} = 1+\frac{1}{x} = 1+\frac{1}{1+\frac{1}{y}}.$$
$$\frac{\sqrt{3}+1}{2} = 1+\frac{1}{y} \to \frac{\sqrt{3}+1}{2}-1 = \frac{1}{y} \to \frac{\sqrt{3}-1}{2} = \frac{1}{y}.$$
$$y = \frac{2}{\sqrt{3}-1}\frac{\sqrt{3}+1}{\sqrt{3}+1} = \frac{2(\sqrt{3}+1)}{2} = 1+\sqrt{3}.$$

The next step in constructing the continued fraction expansion is

$$\sqrt{3} = 1+\frac{1}{x} = 1+\frac{1}{1+\frac{1}{y}} = 1+\frac{1}{1+\frac{1}{1+\sqrt{3}}}. \tag{Ans.2}$$

And now the red part can be replaced with the right-hand side of Equation (Ans.2), and the process repeated ad infinitum. The infinite continued fraction expansion of this irrational is therefore $\sqrt{3} = [1; 1, 2, 1, 2, 1, 2, \ldots]$.

2.25. The Mathematica code below shows the unusual, regular sequence of coefficients of the continued fraction of e. A proof of this (too advanced for this book) can be found in [Cohn 06].

```
In[4]:= ContinuedFraction[E,100]
Out[4]= {2,1,2,1,1,4,1,1,6,1,1,8,1,1,10,1,1,12,1,1,14,1,1,16,1,1,18,1,1,20,1,1,
22,1,1,24,1,1,26,1,1,28,1,1,30,1,1,32,1,1,34,1,1,36,1,1,38,1,1,40,1,1,42,1,1,44,
1,1,46,1,1,48,1,1,50,1,1,52,1,1,54,1,1,56,1,1,58,1,1,60,1,1,62,1,1,64,1,1,66,1}
```

2.26. The fact that the left-hand sides of the two continued fractions are different, suggests that we should consider them (the left-hand sides) an unknown x and write the two continued fractions as the single quadratic equation

$$x = \frac{2}{3-x}.$$

This equation can be written as $(x-1)(x-2) = 0$, which produces the two solutions $x = 1$ and $x = 2$. The right-hand sides of the continued fractions can converge to either 1 or 2 depending on which point they are chopped off. If they are chopped as shown in the exercise, after any of the $(3-2)$'s, then they always equal 2. If they are chopped after any of the 3's, then they converge to 1. The following Mathematica lines show typical outputs in the two cases.

```
2./(3-2/(3-2/(3-2/(3-2/(3-2/(3))))))
0.992126
```

```
2./(3-2/(3-2/(3-2/(3-2/(3-2/(3-2))))))
2.
```

2.27. This is straightforward and results in the similar expression

$$y = \pm\sqrt{\frac{-a+\sqrt{a^2+b^2}}{2}}.$$

2.28. We start with the four roots for $n = 4$. If α is such a root, then $\alpha^4 = 1$, implying that $\alpha^8 = 1^2 = 1$, making α one of the eight roots for $n = 8$. Four more roots are needed, and we guess that $\sqrt{i}$ is one of them. To prove this, we observe that if $\alpha = \sqrt{i}$ then $\alpha^2 = i$ and $\alpha^8 = i^4 = 1$, but what number is $\sqrt{i}$? If we write $\sqrt{i} = a + bi$ and square both sides, we get $i = a^2 + 2abi - b^2$, which implies that $a^2 - b^2 = 0$ and $2ab = 1$. Thus

$$\sqrt{i} = \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}. \quad \text{The other roots are } \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2},\ -\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}, \text{ and } -\frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}.$$

2.29. Figure Ans.3 (reprinted under attribution license from its author, Wiki user Cmglee) shows roots 2 through 6 of a general complex number $Z = (r, \theta)$.

2.30. We start with $(i^3)^{\frac{1}{4}}$. The cartesian coordinates of $-i$ in the Argand diagram are $(0, -1)$, which is why its polar coordinates are $(1, 3\pi/2)$ or, in general $(1, 3\pi/2 + 2n\pi)$ for nonnegative integers n. We therefore write

$$(i^3)^{\frac{1}{4}} = (-i)^{\frac{1}{4}} = \left[e^{i\frac{4n+3}{2}\pi}\right]^{\frac{1}{4}} = \left[e^{i\frac{4n+3}{8}\pi}\right].$$

For $n = 0, 1, 2$, and 3, this yields $e^{i\frac{3\pi}{8}}$, $e^{i\frac{7\pi}{8}}$, $e^{i\frac{11\pi}{8}}$, and $e^{i\frac{15\pi}{8}}$.

We now swap the exponents and compute $(i^{\frac{1}{4}})^3$ using polar coordinates as before, to obtain the results $(e^{i\frac{\pi}{8}})^3 = e^{i\frac{3\pi}{8}}$, $(e^{i\frac{5\pi}{8}})^3 = e^{i\frac{15\pi}{8}}$, $(e^{i\frac{9\pi}{8}})^3 = e^{i\frac{27\pi}{8}}$, and $(e^{i\frac{13\pi}{8}})^3 = e^{i\frac{39\pi}{8}}$. Two of these results equal two parts of the previous results, and the relations $27\pi/8 - 2\pi = 11\pi/8$ and $39\pi/8 - 4\pi = 7\pi/8$ imply that the remaining two results equal the other two parts of the previous results. Thus, for the fractional power 3/4, the order of the exponents does not matter.

2.31. (a). The derivation of $\sqrt[3]{-1}$ is straightforward. Let $a^3 = -1$, then $a^3 + 1 = 0$. We add and subtract both a^2 and a to obtain $a^3 + a^2 - a^2 + a - a + 1 = 0$ which simplifies to $(a + 1)(a^2 - a + 1) = 0$. Hence, one cube root of i is the real number -1 and the other two are the solutions of the quadratic equation $a^2 - a + 1 = 0$, which are $1/2 \pm (\sqrt{3}/2)i \approx 1/2 \pm 0.866i$.

The three cube roots can also be plotted on the Argand diagram as the complex powers $e^{i\pi/3+i2\pi n/3}$ for $n = 0$, 1, and 2. This is illustrated in Figure Ans.4 together with

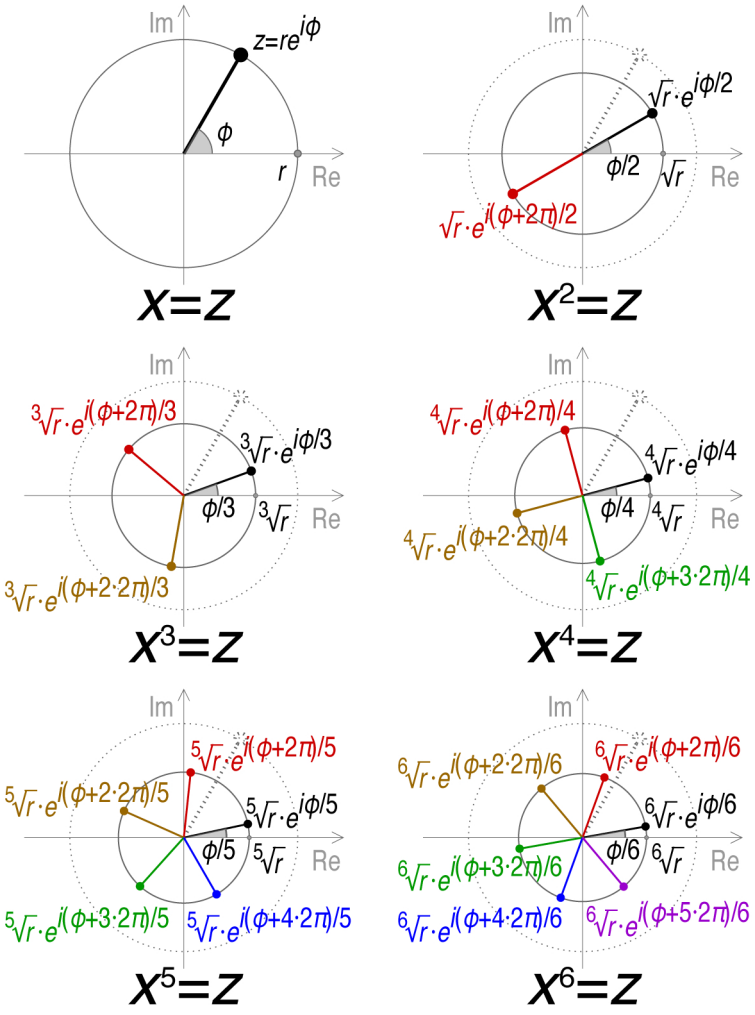


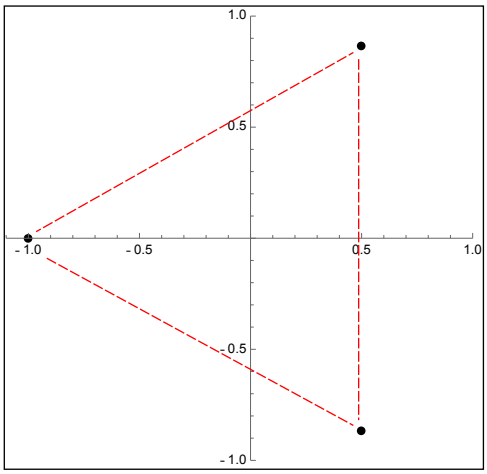
Figure Ans.3: Roots 2 Through 6 of a Complex Number Z.
Reprinted with permission from Wiki user Cmglee.

the Mathematica code. Note that the three roots form the vertices of an equilateral triangle (shown in dashed red lines).

(b). The equation $x^3 = 8$ can be manipulated to produce

$$\begin{aligned} x^3 - 8 &= 0, \\ x^3 - 2^3 &= 0, \\ (x-2)(x^2 + 2x + 2^2) &= 0, \end{aligned}$$

which is easily solved to yield $x = 2$ and $x = -1 \pm i\sqrt{3}$.




```
cr=(1/2)+I Sqrt[3.]/2
cr^3
ComplexListPlot[{{E^(Pi I /3)},{E^(Pi I /3+2Pi I /3)},{E^(Pi I /3+2Pi I 2/3)}},
 PlotRange->{{-1.1,1},{-1,1}}]

0.5 + 0.866025 I
-1. + 1.11022*10^-16 I
```

Figure Ans.4: Three Cube Roots of i.

2.32. An imaginary number has the form ai where a is real. The sum of two such numbers is therefore $ai + bi = (a + b)i$, and this can be real only if it is zero, i.e., when $a = -b$.

2.33. An Internet search has yielded the unpronounceable term trigintaduonions. The Wikipedia article says:

"The trigintaduonions, also known as the 32-ions, 32-nions, 25-nions, or sometimes pathions $\mathbb{P}$, form a 32-dimensional noncommutative, and nonassociative algebra over the real numbers, usually represented by the capital letter T, boldface **T** or blackboard bold $\mathbb{T}$.

The word trigintaduonion is derived from Latin triginta (thirty) + duo (two) + the suffix -nion, which is used for hypercomplex number systems."

(The term pathion may be a reference to the 32 paths of wisdom, found in certain religious writings.)

2.34. Today, we believe that Rene Descartes was one of the first, perhaps *the* first, who coined the term "imaginary" in his book *La Geometrie*, where he wrote:

> "For the rest, neither the false nor the true roots are always real, sometimes they are only imaginary, that is to say one may imagine as many as I said in each equation, but sometimes there exists no quantity corresponding to those one imagines."

To Descartes, imaginary numbers were an alternate method of solving "what-if" types of questions. He thought of them as imagined solutions instead of true solutions. This seemed reasonable to mathematicians of his day. They accepted Descartes' thinking and point of view, and the term "imaginary" stuck.

Then came Gauss, convincing his contemporaries that imaginary quantities make sense by showing that they can be visualized:

> "That this subject [imaginary numbers] has hitherto been surrounded by mysterious obscurity, is to be attributed largely to an ill adapted notation. If, for example, $+1$, -1, and the square root of -1 had been called direct, inverse and lateral units, instead of positive, negative and imaginary (or even impossible), such an obscurity would have been out of the question."

2.35. A simple example is the sequence of three 90° rotations: (1) about the x-axis, (2) about the z-axis, and (3) about the y-axis. A point $(0, 1, 0)$ may find itself, by such a sequence of rotations, first on the z-axis, at $(0, 0, 1)$, then turning on that axis without actually moving, and finally moved by the third rotation to the x-axis, at either $(1, 0, 0)$ or $(-1, 0, 0)$.

However, when the order of the rotations is changed to z, x, and y, point $(0, 1, 0)$ would move to $(1, 0, 0)$, then to itself, and finally to $(0, 0, 1)$.

2.36. Yes, there are two and only two such extensions, octonions and sedenions.

Octonions are a nonassociative extension of the quaternions. They form an eight-dimensional normed division algebra over the real numbers. Octonions were defined and developed in 1843 by John T. Graves, an associate of William Hamilton, who referred to them as octaves.

Sedenions form a 16-dimensional algebra over the real numbers obtained by applying the Cayley-Dickson construction to the octonions. Like octonions, multiplication of sedenions is neither commutative nor associative.

2.37. Using the rule for quaternion multiplication and the three trigonometric identities

$$\cos\theta = \cos^2\tfrac{\theta}{2} - \sin^2\tfrac{\theta}{2}, \quad \sin\theta = 2\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}, \quad \text{and} \quad \cos\theta = 1 - 2\sin^2\tfrac{\theta}{2},$$

we can write

$$\begin{aligned}
\mathbf{q}\cdot[0,\mathbf{P}]\cdot\mathbf{q}^{-1} &= \left[\cos\tfrac{\theta}{2}, \mathbf{u}\sin\tfrac{\theta}{2}\right]\cdot[0,\mathbf{P}]\cdot\left[\cos\tfrac{\theta}{2}, -\mathbf{u}\sin\tfrac{\theta}{2}\right] \\
&= \left\{\left[\cos\tfrac{\theta}{2}, \mathbf{u}\sin\tfrac{\theta}{2}\right]\cdot[0,\mathbf{P}]\right\}\cdot\left[\cos\tfrac{\theta}{2}, -\mathbf{u}\sin\tfrac{\theta}{2}\right] \\
&= \left[-\sin\tfrac{\theta}{2}(\mathbf{u}\bullet\mathbf{P}), \cos\tfrac{\theta}{2}\mathbf{P} + \sin\tfrac{\theta}{2}(\mathbf{u}\times\mathbf{P})\right]\cdot\left[\cos\tfrac{\theta}{2}, -\mathbf{u}\sin\tfrac{\theta}{2}\right] \\
&= [-\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}(\mathbf{u}\bullet\mathbf{P}) + \sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}(\mathbf{P}\bullet\mathbf{u}) - \sin^2\tfrac{\theta}{2}(\mathbf{u}\times\mathbf{P})\bullet\mathbf{u}, \\
&\quad \sin^2\tfrac{\theta}{2}(\mathbf{u}\bullet\mathbf{P})\mathbf{u} + \cos^2\tfrac{\theta}{2}\mathbf{P} + \sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}(\mathbf{u}\times\mathbf{P}) \\
&\quad - \sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}(\mathbf{P}\times\mathbf{u}) - \sin^2\tfrac{\theta}{2}(\mathbf{u}\times\mathbf{P})\times\mathbf{u}] \\
&= \left[0, \sin^2\tfrac{\theta}{2}(\mathbf{u}\bullet\mathbf{P})\mathbf{u} + \cos^2\tfrac{\theta}{2}\mathbf{P} + 2\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}(\mathbf{u}\times\mathbf{P})\right. \\
&\quad \left. - \sin^2\tfrac{\theta}{2}(\mathbf{P} - (\mathbf{u}\bullet\mathbf{P})\mathbf{u})\right]
\end{aligned}$$

$$
\begin{aligned}
&= \left[0, 2\sin^2\tfrac{\theta}{2}(\mathbf{u}\bullet\mathbf{P})\mathbf{u} + (\cos^2\tfrac{\theta}{2} - \sin^2\tfrac{\theta}{2})\mathbf{P} + 2\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}(\mathbf{u}\times\mathbf{P})\right] \\
&= [0, (1-\cos\theta)(\mathbf{u}\bullet\mathbf{P})\mathbf{u} + \cos\theta\mathbf{P} + \sin\theta(\mathbf{u}\times\mathbf{P})] \\
&= [0, (\mathbf{u}\bullet\mathbf{P})\mathbf{u} + \cos\theta(\mathbf{P} - (\mathbf{u}\bullet\mathbf{P})\mathbf{u}) + \sin\theta(\mathbf{u}\times\mathbf{P})],
\end{aligned}
$$

that is Equation (8.15).

2.38. We denote our unknown function and its derivative by $y(x)$ and $y'(x)$, respectively, and require that $yy' = 1$. The derivative of any $y^2(x)$ with respect to x, i.e.

$$\frac{dy^2}{dx},$$

is always $2yy'$. We require that $yy' = 1$, which yields

$$\frac{dy^2}{dx} = 2yy' = 2.$$

Integrating produces $y^2(x) = (2x) + \text{Const}$, which implies $y = \pm\sqrt{2x} + C$.

2.39. The following are three old approximations of π.

- The Bible. The biblical verse often cited in discussions about the approximation of π is *1 Kings* 7:23. This verse describes a large bronze basin, referred to as the "Sea," in Solomon's temple, stating it was 10 cubits in diameter and 30 cubits in circumference. This would imply a value of $\pi = 3$.

- The Rhind papyrus. This 3500-year-old document, written by an ancient scribe named Ahmes and purchased in 1858 by the Scottish antiquarian Alexander Henry Rhind, is one of few ancient mathematical documents that miraculously survived and came down to us. Its two parts contain about 60 arithmetic and algebraic problems, where #48 approximates π as $256/81 \approx 3.16$, a very good approximation, especially for the time it was written, about 1551 B.C.

  Figure Ans.5 illustrates Ahmes's derivation of this value. The octagon (the square minus the four triangles) represents the circle. Its radius is $9/2$ and its area is $9\times 9 - 4(3\times 3/2) = 63$, a value whose square root is close to 8, so Ahmes estimated the area of his circle as 64, the area of a square of side 8 units. The modern relation $A = \pi r^2$ can now be used to produce the estimate

$$\pi = \frac{A}{r^2} = \frac{64}{(9/2)^2} = 64\frac{4}{81} = \frac{256}{81} \approx 3.16.$$

- The Indiana pi bill. The following is quoted from Wikipedia page `Indiana_pi_bill`

  > In 1894, Indiana physician Edward J. Goodwin, believed that he had discovered a way of squaring the circle. He proposed a bill to state representative Taylor I. Record, who introduced it in the House under the title 'A Bill for an act introducing a new mathematical truth and offered as a contribution to education to be used only by the State of Indiana free of cost...'.

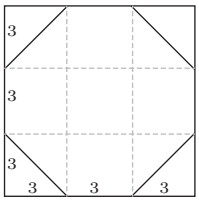


Figure Ans.5: Ancient Egyptian π.

Even though not mentioning π, the bill would have implied that π has the value $4/1.25 = 3.2$, and also that $\sqrt{2} \approx 1.429$. Fortunately, the bill never became law due to the intervention of C. A. Waldo, a professor at Purdue University, who happened to be present in the legislature on the day it went up for a vote.

Imagine the reaction of any aliens happening to visit Indiana on that day.

2.40. Here are a few.

Born: English actor Michael Caine, comedian Billy Crystal.

Died: Physicist and cosmologist Stephen Hawking, philosopher and socialist Karl Marx.

2.41. No, because such an "inner" subsequence would have, in such a case, another sub-subsequence inside it, causing the sequence of π digits to repeat itself, and thereby causing π to be rational.

2.42. We cannot really tell. The best we can claim is that in our world, π is proved transcendental, so in a world where π is non-transcendental, this proof would be invalid and our mathematics would not hold. Would there still be intelligent life in such a world? Is there intelligent life in a world where someone can even ask this question about π? Weird!

2.43. The former is simple. From $(1+i)^2 = 1^2+2i+i^2 = 2i$ we obtain $\sqrt{2i} = 1+i$. We now assume that the $\sqrt{2+i}$ is a complex number, which we write as $a+bi$. If this leads us nowhere, we can always restart by assuming that it is real. From $\sqrt{2+i} = a+bi$ we arrive at $(\sqrt{2+i})^2 = (a+bi)^2$ or $2+i = a^2-b^2+2abi$. Equating the real and imaginary parts yields the equations $a^2 - b^2 = 2$ and $ab = 1/2$.

One way to proceed is to rewrite these in the form $a^2 = b^2 + 2$ and $a^2b^2 = 1/4$, and substitute a^2. This results in $b^2(b^2 + 2) = 1/4$, which can be solved for b^2 and produce a and b.

An alternative is to employ the identity $(x + y)^2 - (x - y)^2 = 4xy$. Applying it to the equations results in $(a^2 + b^2)^2 - (a^2 - b^2)^2 = 4a^2b^2$ or $(a^2 + b^2)^2 - 2 = 4(1/4)$, which simplifies to $(a^2 + b^2)^2 = 5$ or $a^2 + b^2 = \pm\sqrt{5}$. It is obvious that the negative case is meaningless, since the sum $a^2 + b^2$ cannot be negative.

It remains to deal with the positive case. Subtracting $a^2 + b^2 = \sqrt{5}$ and $a^2 - b^2 = 2$ yields

$$a^2 = \frac{\sqrt{5}+2}{2} \quad \text{and} \quad b^2 = \frac{\sqrt{5}-2}{2}.$$

The last step is to visualize the complex number $2+i$ in the Argand diagram. Both its real (2) and imaginary (i) parts are positive, which is why the real (a) and imaginary (b) parts of $\sqrt{2+i}$ must also be positive. The final solution is therefore

$$\sqrt{2+i} = \sqrt{\frac{\sqrt{5}+2}{2}} + \sqrt{\frac{\sqrt{5}-2}{2}}\,i \approx 1.45535 + 0.343561i.$$

2.44. The error is in step 5. It is valid to divide a real number by an imaginary number, but replacing $\sqrt{a}/\sqrt{b}$ with $\sqrt{a/b}$, as in step 2, is valid only if a is non-negative and b is positive. We can also consider the square root a function $f(x)$, but then we have to prove that it satisfies $f(a)/f(b) = f(a/b)$.

2.45. From the discussion of i^i on Page 144 we know that $i = e^{i\frac{\pi}{2}}$ and $i^i = e^{-\frac{\pi}{2}}$, from which we easily deduce that

$$i^{(i^i)} = \left[e^{i\frac{\pi}{2}}\right]^{e^{-\frac{\pi}{2}}} = e^{i\frac{\pi}{2}e^{-\frac{\pi}{2}}} = \cos(\frac{\pi}{2}e^{-\frac{\pi}{2}}) + i\sin(\frac{\pi}{2}e^{-\frac{\pi}{2}}).$$

This expression, which some may consider ugly, is nevertheless correct.

2.46. This is trivial. $1/i = -i$ because $i\cdot(-i) = -i^2 = -(-1) = 1$.

2.47. This is straightforward

$$\begin{aligned}
&\sqrt{i} + \sqrt{-i} \\
&= \frac{1}{\sqrt{2}}\sqrt{2i} + \frac{1}{\sqrt{2}}\sqrt{-2i} \\
&= \frac{1}{\sqrt{2}}\sqrt{1+2i-1} + \frac{1}{\sqrt{2}}\sqrt{1-2i-1} \\
&= \frac{1}{\sqrt{2}}\left[\sqrt{(1+i)^2} + \sqrt{(1-i)^2}\right] \\
&= \frac{1}{\sqrt{2}}\left[\pm(1+i) + \pm(1-i)\right] \\
&= \pm\sqrt{2}.
\end{aligned}$$

2.48. Applying the Gamma function, Equation (10.1), to i, the derivation is straightforward but tedious, and the result is a certain complex number that doesn't seem to

have any special meaning or interest.

$$\begin{aligned}
i! &= \int_0^\infty x^z e^{-x}\,dx \\
&= \int_0^\infty e^{\ln(x^i)} e^{-x}\,dx \\
&= \int_0^\infty e^{i\ln(x)} e^{-x}\,dx \\
&= \int_0^\infty (\cos(\ln x) + i\sin(\ln x)) e^{-x}\,dx \\
&= \int_0^\infty [\cos(\ln x) e^{-x} + i\sin(\ln x) e^{-x}]\,dx \\
&= \int_0^\infty \cos(\ln x) e^{-x}\,dx + i\int_0^\infty \sin(\ln x) e^{-x}\,dx \\
&\approx 0.498 - 0.155i.
\end{aligned}$$

As far as I know, there are no closed form expressions for the last two integrals. They have to be computed numerically. Also, the final result does not provide any insight into the nature of i or its factorial. It is just a number.

2.49. Solution 1 uses the famous Euler equation

$$e^{i\pi} = -1,$$

and proceeds in the following steps:

$$\begin{aligned}
(-5)^x &= 5, \\
(-1)^x 5^x &= 5, \\
e^{i\pi x} 5^x &= 5, \\
\ln\left[e^{i\pi x}\right] + \ln 5^x &= \ln 5, \\
i\pi x \ln e + x \ln 5 &= \ln 5, \\
x &= \frac{\ln 5}{i\pi + \ln 5}.
\end{aligned}$$

The second solution is simpler

$$\begin{aligned}
(-5)^x &= 5, \\
\ln(i^2 \cdot 5)^x &= \ln 5, \\
x\left[\ln(i^2) + \ln 5\right] &= \ln 5, \\
x &= \frac{\ln 5}{2\ln(i) + \ln 5} = \frac{\ln 5}{i\pi + \ln 5}.
\end{aligned}$$

2.50. The answer depends on n mod 4 and can have one of the four values

$$i^{n+1} - i^n = i^n(i-1) = \begin{cases} n \bmod 4 = 1, & i(i-1); \\ n \bmod 4 = 2, & -1(i-1); \\ n \bmod 4 = 3, & -i(i-1); \\ n \bmod 4 = 0, & 1(i-1). \end{cases}$$

2.51. The answer, naturally, is a matter of personal opinion, but we can claim that so many numbers, including integers and nonintegers, have interesting properties, that this feature of ϕ is perhaps not as fundamental as seems at first.

2.52. Draw a square of side a. With the compass, locate the midpoint M of one side. Draw a circle (in black) of radius BM centered at M. Notice that it touches corners A and B. Use this circle to locate point C that is on the line DF. Complete the rectangle by locating the midpoint of AB and drawing a circle (in blue) at that point that will touch corners D and F. Extend segment AB to touch the circle, which will determine point E.

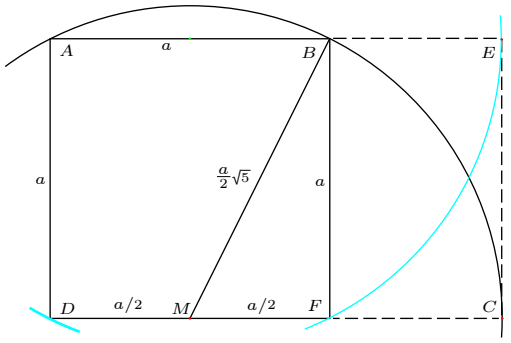


Figure Ans.6: A Golden Rectangle with Ruler and Compass.

The radius of the black circle (segments MB and MC) is

$$\sqrt{\frac{a^2}{2} + a^2} = \frac{a}{2}\sqrt{5},$$

which is why the length of segment DC is

$$\frac{a}{2} + \frac{a}{2}\sqrt{5} = a(1+\sqrt{5})/2 = a\phi,$$

showing that rectangle $ADCE$ is golden.

Figure Ans.7 illustrates several geometric constructions that yield Φ and related constants.

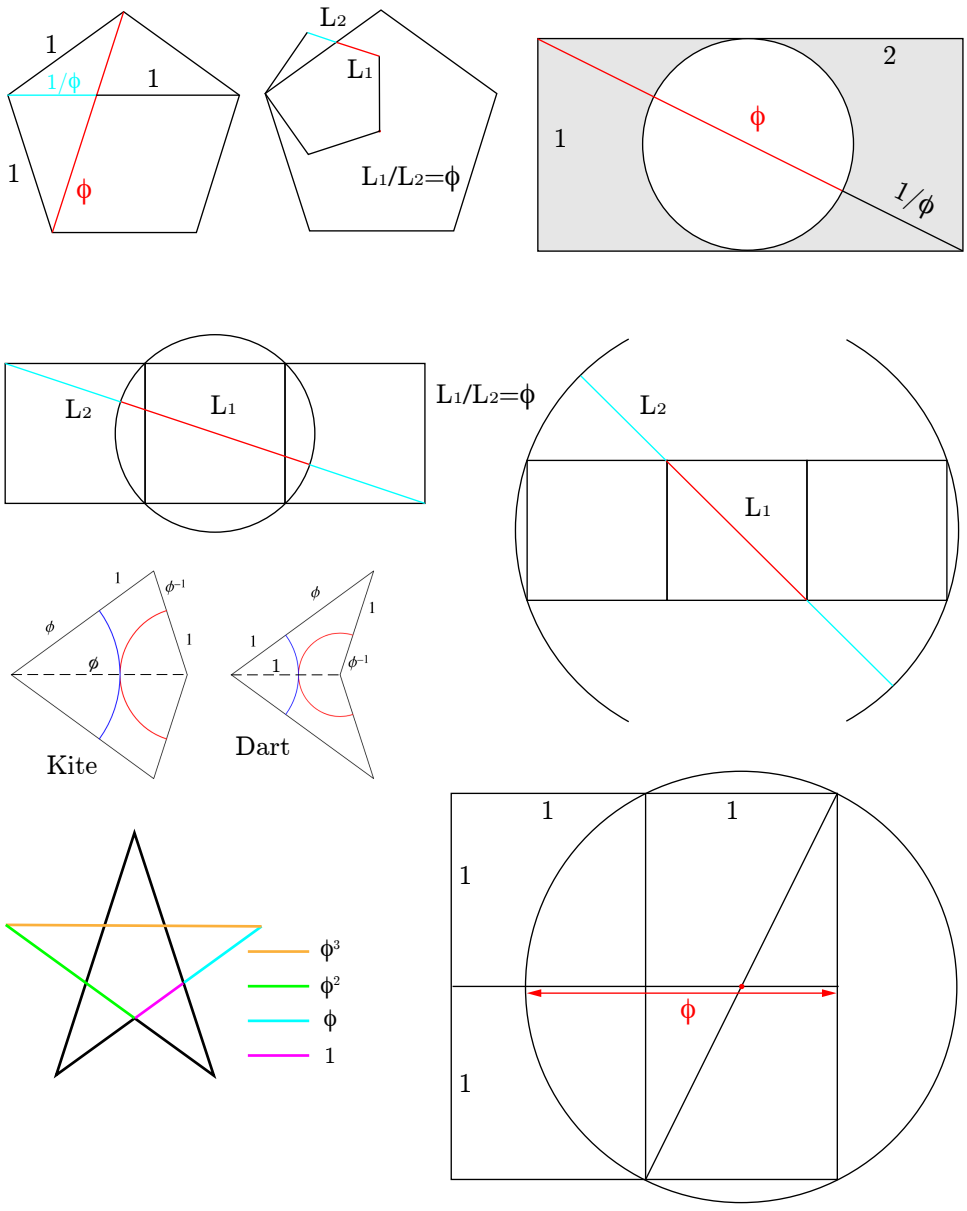


Figure Ans.7: A Golden Zoo of ϕ and L.

2.53. A simple explanation is that the golden ratio is a solution of the quadratic equation

$$x^2 - x - 1 = 0. \tag{Ans.3}$$

Writing this in the form $x^2 = x + 1$, Figure Ans.8 shows that for $x = 0$, the right-hand-side (RHS, the blue straight line $y = x + 1$) is bigger than the LHS (the red parabola), but for $x = 2$, the RHS is smaller. Because the two curves are continuous, there must be a point `A` between $x = 0$ and $x = 2$ where the two curves meet. This point, which is a solution of Equation (Ans.3), is named the golden ratio ϕ. The other solution of this equation is point `B` between $x = -1$, where RHS $<$ LHS and $x = 0$, where RHS $>$ LHS, that we denote by $1/\phi$.

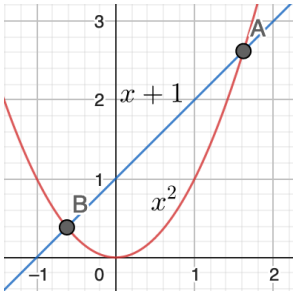


Figure Ans.8: Two Intersecting Curves at ϕ.

Another answer to this tricky question is that the study of quadratic equations shows that the real number $(1 + \sqrt{5})/2$ is the positive solution of Equation (Ans.3).

2.54. The two solutions of the quadratic equation $x^2 - ix + 1 = 0$ are the complex numbers $(i \pm i\sqrt{5})/2$, as can be verified by direct checking.

Similarly, the two solutions of $2x^2 - 2x + 3 = 0$ are the imaginaries $(1 \pm i\sqrt{5})/2$, as can be verified by the single Mathematica statement `Solve[2x^2-2x+3==0,x]`

Also, the fifth roots of unity, Page 109, are related to ϕ, and four of those roots are complex.

2.55. For $n = 2$ the noem is $1 + \sqrt{2}$ as can be derived from Figure 2.47a.

2.56. This is true, but the zero can be included in several places by, for example, writing 03 instead of 3. Alternatively, we can get clever and claim that Sabey's formula is only weakly pandigital.

3.1. Examining the hands, it is clear that the symmetry of this image is not bilateral but rather rotational symmetry about a vertical axis at the center of the figure.

3.2. A common example is the pedal threads in a bicycle, which are different for the left and right sides. The right side pedal has a right-hand thread (removes counterclockwise, installs clockwise). The left side pedal has a left-hand thread (removes clockwise, installs counterclockwise). Similarly, left-hand threads are used in any engine and turbine parts, where the rotation can cause right-handed threads to loosen over time.

There are cases where left-hand threads are used to intentionally create incompatibility as a safety indicator. Gas-supply valves use left-handed threads to differentiate them from the valves that control oxygen.

Used in fiction. In his 1974 book *Rendezvous with Rama*, Arthur C. Clarke says, "there was no reason to suppose that clocks and corkscrews on Rama turned in the same direction as they did on Earth."

3.3. (By chatGPT in October 2023).

No, not every group is a symmetry group. A symmetry group is a specific type of mathematical group that describes the symmetries of a particular object or system. It is a set of transformations that leave an object or system unchanged, such as rotations or reflections. Not all groups have this property, and therefore not all groups are symmetry groups.

3.4. The only subgroups of G are G itself and the trivial group.

3.5. The process is straightforward and the result is $(1, 6, 2, 9, 7, 4)$

3.6. $D_8 = \{(), (1234), (13)(24), (1432), (24), (13), (14), (23)\}$.

3.7. The result, known as the "average die" has faces numbered 2, 3, 3, 4, 4, and 5.

3.8. This is easy. Figure Ans.9 shows how a reflection about a horizontal line that passes through the blue dot results in the green figure, which is rotated about the same dot to produce the cyan figure. Finally, another reflection about a vertical line that passes through the blue dot brings the cyan figure to its original, red orientation and location (in the figure, it is slightly shifted to make it visible).

The point of this exercise is to stress the important fact that a flat, two-dimensional figure does not have front and back. It is infinitely thin, which is why it is easy to visualize two-dimensional reflections and why they are often symmetry transformations.

3.9. The following was generated by ChatGPT in October 2023:

The symmetry group of a cube is known as the octahedral symmetry group and is often denoted as O or O_h. It consists of 48 distinct symmetries, including rotations and reflections, that preserve the shape and position of the cube in three-dimensional space. To represent the multiplication table for this group, we can use a Cayley table.

The 48 symmetries can be categorized into the following:

1. Identity (E): No transformation; the cube remains unchanged.

2. Rotations (Cn): There are 24 rotational symmetries, which include rotations of 90, 180, and 270 degrees around various axes.

3. Reflections (σ_h): There are 11 horizontal mirror reflections that reflect the cube across a plane perpendicular to one of the cube's faces.

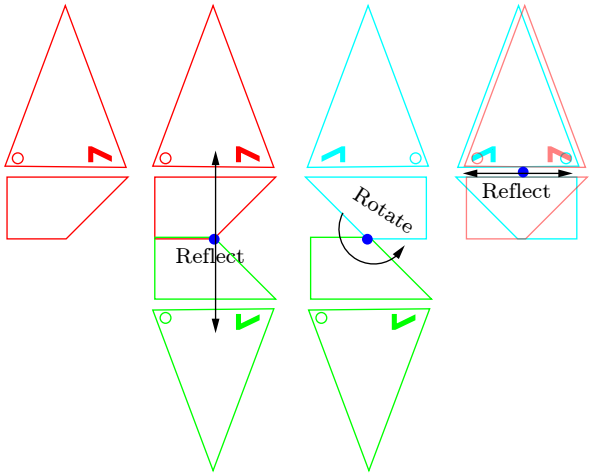


Figure Ans.9: Transformations for Exercise 3.8.

4. Reflections (σ_v): There are 6 vertical mirror reflections that reflect the cube across a plane containing two opposite vertices.

5. Reflections (Sd): There are 6 diagonal mirror reflections that reflect the cube across a plane containing two opposite edges.

6. Rotational reflections (S4): There is 1 symmetry that combines a 90-degree rotation with a reflection across a plane perpendicular to the axis of rotation.

The Cayley table (group multiplication table) of the octahedral symmetry group is quite extensive, as it involves the compositions of all these symmetries. It's typically represented as a 48×48 table, which would be too large to display here in its entirety. However, I can provide a simplified version of the table that shows the composition of a few selected symmetries. For the full table, you may refer to group theory textbooks or online resources.

Figure Ans.10 is a simplified portion of the Cayley table for the octahedral symmetry group:

3.10. This is because the five faces of the square-based pyramid are not equal. The base is a square while the four sides are triangular. In contrast, the four faces of the triangular-based pyramid are identical, which endows it with more symmetry. Similarly, the octahedron, which consists of two square-based pyramids connected at their bases, also features higher symmetry (its symmetry is isomorphic to that of the cube).

3.11. The following rephrases and extends reference [Reid 04], a detailed discussion of symmetries in brick walls, which also considers different brick patterns and how they affect the symmetries of such a wall. Figure Ans.11 shows only the brick pattern known as "running bond," which is the most common. Each brick has four sides which can be viewed as six edges (in dashed blue). The following symmetries are easy to visualize:

- Translation. The black arrows show how an infinite wall can be translated to itself horizontally, vertically, and diagonally.

```
css                                                                    Copy code

 * | E  C2x C2y C2z σh1 σh2 σh3 σh4 σh5 σh6 σv1 σv2 σv3 σv4 σv5 σv6 σd1 σd2
-----------------------------------------------------------------------------
E  | E  C2x C2y C2z σh1 σh2 σh3 σh4 σh5 σh6 σv1 σv2 σv3 σv4 σv5 σv6 σd1 σd2
C2x| C2x E   σv1 σv3 σh1 σh3 σd1 σd2 C2y C2z σv5 σv6 σv2 σv4 σh5 σh6 σh2 σh4
C2y| C2y σv2 E   σv4 σh2 σh4 σd2 σd3 C2x C2z σv6 σv1 σv3 σv5 σh6 σh1 σh3 σh5
C2z| C2z σv3 σv4 E   σh3 σh4 σd3 σd1 σd2 σd4 C2x C2y σv4 σv6 σv5 σv2 σh4 σh6
σh1| σh1 σh2 σh3 σh4 E   C2x C2y C2z S4 σv1 σv2 σv3 σv4 σv5 σv6 σd1 σd2 σd3
σh2| σh2 σh1 σh4 σh3 C2x E   S4 C2z C2y σv2 σv1 σv4 σv3 σv6 σv5 σd2 σd1 σd4
σh3| σh3 σh4 σh1 σh2 C2y S4 E   C2x C2z σv3 σv4 σv1 σv2 σv5 σv6 σd3 σd1 σd2
σh4| σh4 σh3 σh2 σh1 C2z C2x C2y E   S4 σv4 σv3 σv2 σv1 σv6 σv5 σd4 σd3 σd2
σh5| σh5 σh6 σh6 σh5 σd1 σd2 σd3 σd4 σd5 σd6 E   S4 C2x C2y C2z σv1 σv2 σv3
σh6| σh6 σh5 σh5 σh6 σd2 σd1 σd4 σd3 σd6 σd5 S4 E   C2z C2x C2y σv2 σv1 σv4
σv1| σv1 σv2 σv3 σv4 σv5 σv6 σv1 σv2 σv3 σv4 σh1 σh2 σh3 σh4 E   C2x C2y C2z
σv2| σv2 σv1 σv4 σv3 σv6 σv5 σv2 σv1 σv4 σv3 σh2 σh1 σh4 σh3 C2x E   C2z σd2
σv3| σv3 σv4 σv1 σv2 σv5 σv6 σv3 σv4 σv1 σv2 σh3 σh4 σh1 σh2 C2y C2z E   C2x
σv4| σv4 σv3 σv2 σv1 σv6 σv5 σv4 σv3 σv2 σv1 σh4 σh3 σh2 σh1 C2z C2x C2y E
σv5| σv5 σv6 σv6 σv5 σd3 σd4 σd1 σd2 σd5 σd6 σv5 σv6 σv1 σv2 σv3 σv4 E   C2x
σv6| σv6 σv5 σv5 σ
```

Figure Ans.10: Part of the Octahedral Symmetry Group O_h.

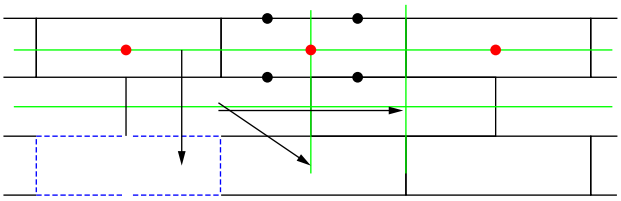


Figure Ans.11: Symmetries of a Brick Wall.

- Reflection. The green lines can be viewed as mirrors about which the wall can be reflected. The horizontal mirrors are located in the middle of each brick layer and the vertical mirrors are located between bricks and at the center of a brick, between its vertical edges.

- Rotation. The circles indicate some of the rotation points. The black circles are located between the horizontal edges of a brick and the red circles are at the centers of the bricks.

- Gyration. The black circles are also gyration points. They are centers of rotation, but are not on any line of reflection.

3.12. The string `OXIHHIXO` has this signature, but only in certain typefaces.

3.13. Two simple patterns are illustrated by Figure Ans.12. Both feature rotation symmetries but no gyrations.

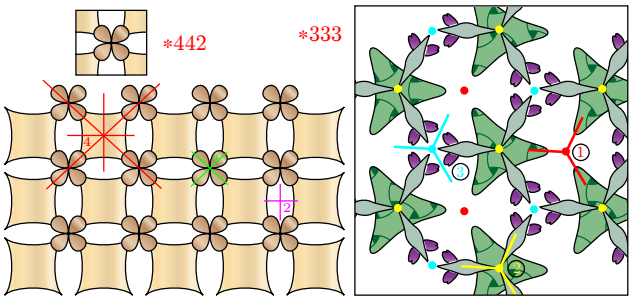


Figure Ans.12: Patterns with $*442$ and $*333$ Symmetries.

3.14. Such a pattern is shown in Figure Ans.13 where four equivalent gyration centers are shown in blue.

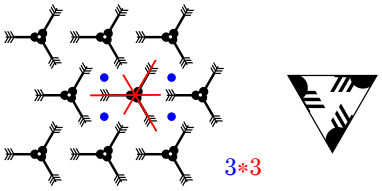


Figure Ans.13: A Pattern with Signature $3*3$.

3.15. Figure Ans.14 shows such a pattern. Notice that in spite of its rich symmetry, this pattern has no reflection axes.

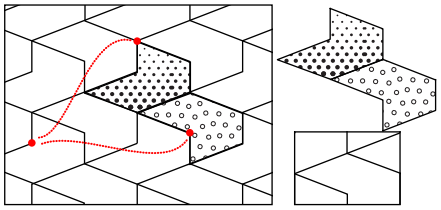


Figure Ans.14: An Otherwise Asymmetric Pattern with Signature $\times\times$.

3.16. A point (x, y) on a circle with radius R satisfies $x^2+y^2 = R^2$ or $(x/R)^2+(y/R)^2 = 1$. The transformed point (x^*, y^*) on an ellipse should satisfy $(x/a)^2 + (y/b)^2 = 1$. It is easy to guess that the transformation rule is $x^* = ax/R$, $y^* = by/R$, but this can also be proved as follows: The general scaling transformation is $x^* = k_1x$, $y^* = k_2y$. For the transformed point to be on an ellipse, it should satisfy $(k_1x/a)^2 + (k_2y/b)^2 = 1$, which can be simplified to $k_1^2b^2x^2 + k_2^2a^2y^2 = a^2b^2$. Substituting $y^2 = R^2 - x^2$ yields

$$(k_1^2b^2 - k_2^2a^2)x^2 = a^2b^2 - k_2^2a^2R^2.$$

This equation must hold for every value of x and this is possible only if $k_1^2b^2 - k_2^2a^2 = 0$ and $a^2b^2 - k_2^2a^2R^2 = 0$. Solving these equations yields $k_1 = a/R$ and $k_2 = b/R$.

3.17. The transformation can be written $(x, y) \to (x, -x + y)$, so $(1, 0) \to (1, -1)$, $(3, 0) \to (3, -3)$, $(1, 1) \to (1, 0)$, and $(3, 1) \to (3, -2)$. The original rectangle is therefore transformed into a parallelogram.

3.18. From $\cos 45^\circ = 0.7071$ and $\tan 45^\circ = 1$, we get the 45°-rotation matrix as the product:

$$\begin{pmatrix} 0.7071 & 0 \\ 0 & 0.7071 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

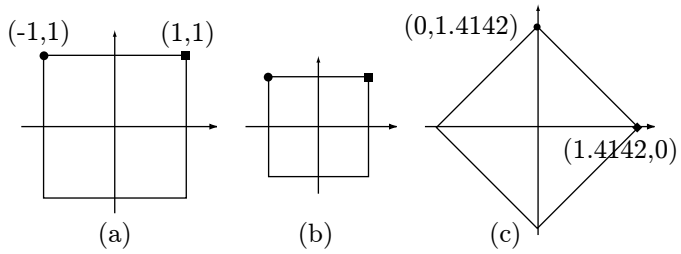


Figure Ans.15: A 45° Rotation as Scaling and Shearing.

Figure Ans.15 shows how a 2×2 square centered on the origin (Figure Ans.15a) is first shrunk to about 70% of its original size (Figure Ans.15b), then sheared by the second matrix according to $(x^*, y^*) = (x+y, -x+y)$, and becomes the rotated diamond shape of Figure Ans.15c. Direct calculations show how the two original corners $(-1, 1)$ and $(1, 1)$ are transformed to $(0, 1.4142)$ and $(1.4142, 0)$, respectively.

3.19. Figure 3.45 gives the polar coordinates $\mathbf{P} = (r, \alpha)$ and $\mathbf{P}^* = (r, \phi) = (r, \alpha - \theta)$. We are looking for a matrix $\mathbf{T} = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ such that $\mathbf{P}^* = \mathbf{PT}$ and it does not take much to figure out that

$$\mathbf{T} = \begin{pmatrix} 1 & -\theta/r \\ 0 & 1 \end{pmatrix}.$$

3.20. A reflection about the x axis transforms a point (x, y) to a point $(x, -y)$. A reflection about $y = -x$ similarly transforms a point (x, y) to a point $(-y, -x)$ (this is matrix $\mathbf{T}_3$ of Equation (3.3)). The combination of these two transformations thus transforms (x, y) to $(y, -x)$, which is another form of the negate and exchange rule, corresponding to a 90° clockwise rotation about the origin. This rotation can also be expressed by the matrix

$$\begin{pmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

3.21. The determinant of this matrix equals

$$\left(\frac{1-t^2}{1+t^2}\right)^2 - \frac{-4t^2}{(1+t^2)^2} = \frac{(1-t^2)^2 + 4t^2}{(1+t^2)^2} = +1,$$

which shows that it generates pure rotation. Also, if we denote this matrix by

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

it is easy to see that $a_{11} = a_{22}$, $a_{12} = -a_{21}$, $a_{11}^2 + a_{12}^2 = 1$, and $a_{21}^2 + a_{22}^2 = 1$. These properties are all satisfied by a rotation matrix.

3.22. The determinant of this matrix is

$$\left(\frac{a}{A}\right)^2 - \frac{b}{A}\left(-\frac{b}{A}\right) = \frac{a^2 + b^2}{A^2}.$$

It equals 1 for $A = \pm\sqrt{a^2 + b^2}$ but cannot equal -1 since it is the quotient of the two non-negative numbers $a^2 + b^2$ and A^2. We, consequently, conclude that this matrix represents pure rotation. An example is $a = b = 1$, which produces $A = \pm\sqrt{2} \approx \pm 1.414$. The rotation matrices for this case are

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{pmatrix},$$
$$\begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} -0.7071 & -0.7071 \\ 0.7071 & -0.7071 \end{pmatrix},$$

and they correspond to 45° rotations about the origin.

3.23. Two reflections of an object about parallel lines are equivalent to a translation, and a translation can be considered a rotation about a point at infinity. But two parallel lines can also be considered as meeting at infinity. Two reflections about the same line return an object to its original position, and are therefore equivalent to a rotation of zero degrees.

"Lines that are parallel
meet at infinity!"
Euclid repeatedly,
heatedly,
  urged
until he died.
and so reached that vicinity:
in it he
found that the damned things
  diverged.
  —Piet Hein (dedicated to Martin Gardner).

3.24. By the definition of matrix A we have

$$A \stackrel{\text{def}}{=} I - R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} 1-\cos\theta & -\sin\theta \\ \sin\theta & 1-\cos\theta \end{pmatrix},$$

so its determinant equals $(1-\cos\theta)^2 + \sin^2\theta = 2 - 2\cos\theta$, and this can be zero only when $\cos\theta = 1$, which happens when θ is zero or a multiple of 2π, but in such cases, a rotation by θ reduces to no rotation.

3.25. Figure Ans.16 lists the steps in preparing the tessellation of the well-known Escher horse.

3.26. underfund, underrefund, and underwound (as in DNA supercoiling) seem to be the only ones. In the book *Labyrinths of Reason*, Sherlock Holmes tries the word `UNDACHSHUND`, but erases it later.

3.27. Only, maybe, perhaps, possibly.

3.28. This mystery is easily solved if we look at it from the point of view of the driver.

3.29. Here are some in English. Skimitry, symmetrie, altimetry, asymmetry, c-symmetry, cp-symmetry, dissymmetry, dosimetry, perimetry, t-symmetry.

3.30. This is trivial because in any base, the infinite sequence of numbers written (in that base) as 101, 1001, 10001, 100001, ... consists of palindromic numbers.

3.31. The divisibility test of 11 is found on Page 61. Scan the digits from right to left and perform an alternating sum, i.e., $a_0 - a_1 + a_2 - a_3 + \ldots$. If the result is divisible by 11, then the entire number is. A little thinking shows that applying this test to a palindrome with an even number of digits results in a zero.

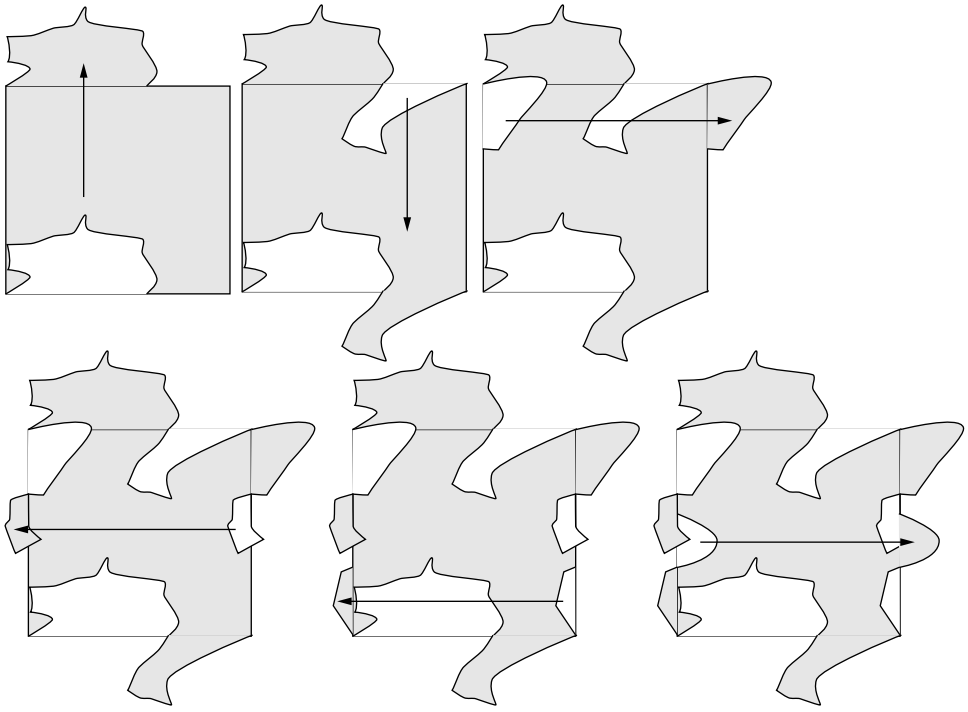


Figure Ans.16: Steps in Constructing the Horse From a Rectangle.

4.1. This is another unexpected effect of repeating a process to infinity. The fragmentation process preserves the circumference of the square, but shrinks its area in each step. At the limit, the area of the fragmented square is π, but its circumference is still 4. It has become an infinitely fragmented square, a fractal (Section 8.1). Its circumference is infinitely fragmented and is nondifferentiable at every real point. The final square looks like a circle, but its circumference is not smoothly curved like that of a circle.

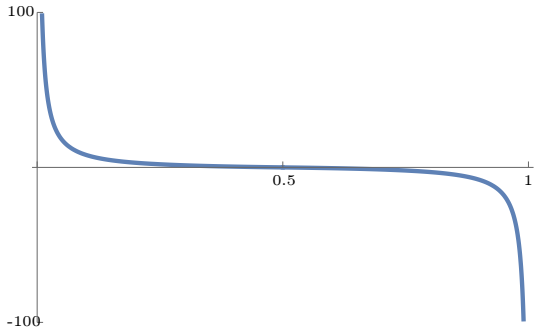
4.2. The specially-crafted list on the left below results in the special number $S = 0.999\ldots$ that is also included in the list.

$$\begin{array}{ll}
1 \to R_1 = 0.99999999\ldots & 1 \to R_1 = 0.3572690\ldots \\
2 \to R_2 = 0.x9xxxx\ldots & 2 \to R_2 = 0.7096418\ldots \\
3 \to R_3 = 0.xx9xxx\ldots & 3 \to R_3 = 0.5381905\ldots \\
4 \to R_4 = 0.xxx9xx\ldots & 4 \to R_4 = 0.1033687\ldots \\
5 \to R_5 = 0.xxxx9x\ldots & 5 \to R_5 = 0.0036731\ldots \\
6 \to R_6 = 0.xxxxx9\ldots & 6 \to R_6 = 0.1958208\ldots
\end{array}$$

A possible way to avoid this special case is to construct the diagonal as follows: If the ith digit of R_i is 7, generate a zero in position i of S, otherwise generate a 7. The result of the list on the right would, in such a case, be $S = 0.0000700\ldots$, a number that cannot be on the list.

4.3. The following bijection, Figure Ans.17, transforms the interval $(0, 1)$ to the entire real number line

$$f(x) = \begin{cases} \frac{1}{x} - 2, & x \in (0, \frac{1}{2}]; \\ 2 - \frac{1}{1-x}, & x \in (\frac{1}{2}, 1). \end{cases}$$




```
f[x_] := If[0 < x <= 1/2, (1/x) - 2, 2 - 1/(1 - x)];
Plot[f[x], {x, 0.01, 0.99}, PlotRange -> 100,
 Ticks -> {{0, 0.5, 1}, {-100, 100}}]
```

Figure Ans.17: Interval $(0, 1)$ transformed.

4.4. We start with a bijection from the set of integers to that of the natural numbers (nonnegative integers). These sets are infinite and it has been known that if there is a bijection between two infinite sets, then there are infinitely many such bijections. Here are two.

Perhaps the simplest bijection is $(0, 1, -1, 2, -2, 3, -3, \ldots) \to (0, 1, 2, 3, 4, 5, 6, \ldots)$. Its inverse is $(0, 1, 2, 3, 4, 5, 6, \ldots) \to (0, 1, -1, 2, -2, 3, -3, \ldots)$.

Another simple bijection is to map the nonnegative integers to the even positive integers ($n \to 2n$), and map all the negative integers to the odd positive integers by $n \to -(2n + 1)$.

We can now denote any such bijection by $a_i \to b_i$, where a_i is an integer and b_i is a nonnegative integer.

An algebraic number is the solution of a polynomial equation with integer coefficients. Such equations have the form $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = 0$. We now construct a transformation F from such an equation to the natural number A defined as the product $2^{b_0}3^{b_1}5^{b_3} \cdots p_i^{b_n}$, where p_i is the ith prime number and the various b_i are the images of a_i under any bijection from the integers to the naturals. The number A is positive because all the exponents b_i are nonnegative. Also, A can equal 1 only if all the b_i are zeros, but in such a case we can choose a bijection where the integer 0 is mapped to the natural number 0. This would imply that all the a_i coefficients are zero and the polynomial equation does not exist.

The fundamental theorem of arithmetic states that any natural number greater than 1 can be uniquely represented as the product of prime numbers. This implies that

every polynomial equation with integer coefficients can be transformed by F to a unique positive integer A, and any positive integer is the transform of such an equation. Thus, transformation F is a bijection, which implies that the sets of natural numbers and of polynomial equations have the same cardinality, and the cardinality of the algebraic numbers is therefore $\aleph_0$.

4.5. This is obtained by differentiating x^x and equating the derivative to zero. We start by considering the two functions $u(x)$ and $v(x)$ and differentiating

$$\frac{d(u(x)^{v(x)})}{dx} = v(x)u(x)^{v(x)-1}u'(x) + \log_e(u(x))u(x)^{v(x)}v'(x).$$

In the special case where $u(x) = v(x) = x$, it is straightforward to obtain the derivative

$$\frac{d(x^x)}{dx} = xx^{x-1} + \log_e(x)x^x = x^x(1 + \log_e(x)).$$

Finally, the Mathematica statement `Solve[x^x(1+Log[e,x])==0, x]` produces $1/e$.

4.6. The first approach is straightforward and employs Euler's equation $e^{i\pi} = 1$.

$$\begin{aligned}
(-1 \cdot 2)^x &= 2, \\
(-1)^x 2^x &= 2, \\
e^{i\pi x} 2^x &= 2, \\
\ln\left(e^{i\pi x}\right) + \ln 2^x &= \ln 2, \\
x(i\pi + \ln 2) &= \ln 2, \\
x &= \frac{\ln 2}{i\pi + \ln 2}.
\end{aligned}$$

The second method is straightforward and is based on a logarithm to any base

$$\begin{aligned}
(-1 \cdot 2)^x &= 2, \\
(2i^2)^x &= 2, \\
x \log(2i^2) &= \log 2, \\
x(2 \log i + \log 2) &= \log 2, \\
x &= \frac{\log 2}{2 \log i + \log 2}.
\end{aligned}$$

5.1. In three days, a chicken and a half will lay three eggs. Thus, a chicken and a half lays one egg a day, and one chicken lays 2/3 an egg in one day. Easy, but what about the following: If a chicken and a half lays an egg and a half in a day and a half then how long would it take a cricket with a wooden leg to kick all the seeds out of a dill pickle?



Figure Ans.18: Image by MS Copilot, Sept. 2025.

5.2. Multiplying both sides by $1-x$ produces $x-1=1-x$ or $x=1$, but a simple check verifies that this results in $0/0=1$, a contradiction.

While we are at it, the similar equation

$$\frac{x-1}{1-x}=-1,$$

implies $x-1=-(1-x)=x-1$, which is meaningless.

5.3. Because the area of a square of side x is x^2 and this is also the dominant term in a degree-2 equation. Similarly for the volume of a cube.

5.4. One reason is that mathematicians are curious. Also, history shows that it is impossible to tell which mathematical and scientific ideas and theories would prove useful in future. An answer that we often hear from mountain climbers is "because they are there."

5.5.

We define $x=1-\sqrt{5}$ to obtain $1-x=\sqrt{5}$ and $(1-x)^2=1-2x+x^2=5$ or

$$x^2=2x+4. \tag{Ans.4}$$

Equation (Ans.4) is now used in the following derivation of x^4

$$\begin{aligned} x^4=(x^2)^2=(2x+4)^2&=16+16x+4x^2\\ &\overset{\text{(Ans.4)}}{=}16+16x+4(2x+4)=32+24x, \end{aligned}$$

as well as in the derivation of x^8

$$\begin{aligned} x^8=(x^4)^2=(32+24x)^2&=32^2+2\cdot 32\cdot 24x+24^2x^2\\ &=1024+1536x+576x^2\overset{\text{(Ans.4)}}{=}1024+1536x+576(2x+4)\\ &=3328+2688x=3328+2688(1-\sqrt{5})=6016-2688\sqrt{5}. \end{aligned}$$

A calculator shows that the last expression equals approximately 5.44928.

This technique can be applied to similar simplification problems. Try, for example, $((1+\sqrt{5})/2)^{12}$.

5.6. Yes, except zero. Even irrational and transcendental numbers, and even imaginary and complex numbers, can serve as bases of a number system. The details of this are tedious, and the interested reader is referred to [any.number 23]. However, in a numbering system based on an irrational number, such as π, e, or $\sqrt{2}$, almost every integer would become an irrational; resulting in infinitely-long computations.

5.7. This is true for numbers to base 7, because the infinite sum

$$\sum_{i=1}^{\infty} \frac{6}{7^i},$$

converges to 1.

5.8. A quick search yields the following: buttercups have 5 petals; lilies and iris have 3 petals; some delphiniums have 8; corn marigolds have 13 petals; some asters have 21 whereas daisies can be found with 34, 55 or even 89 petals.

5.9. The rational $(a+xb)/(a+b)$ is closer to $\sqrt{x}$ than is a/b. This is referred to as the Bhaskara-Brouncker algorithm and it shows that there are infinitely many rationals that get closer and closer to $\sqrt{x}$.

5.10. The rule of the first sequence is to increase the number of holes in the glyphs of the digits that make up the current element. It relies on the fact that in many fonts, the digit 4 has a hole in its glyph. In a font where the 4 is open, the 6 can instead be used. Thus, 4888 has a total of 7 holes, one more than in its predecessor, 888. This is sequence oeis A249572 (holey numbers).

The second sequence is simply the multiples—6, 12, 18, 24, 30, 36, ...—where the commas separating the elements have been moved one digit to the right. This is sequence oeis A087409 (sixers).

The third sequence (very tricky) is based on Roman numerals. To determine the nth element, we first write the integer n in English, and then concatenate all occurrences of the letters I, V, X, L, C, D, and M (which stand for 1, 5, 10, 50, 100, 500, and 1,000, respectively), and compute the Roman numeric value of the resulting string of letters. If the string is invalid (as, for example, in IDIV), then the corresponding element is defined as 0. This is sequence oeis A002904 (think like a Roman).

The fourth sequence (oeis A006933) is referred to as the Eban sequence (for `E`-is-banned). It consists of the positive integers whose English names do not contain the letter `E`. It is surprising that the integers 67 through 1999 have one or more `E`s in their names.

Finally, sequence (oeis A006567) which is dubbed `Emirps`, consists of the non-palindromic primes.

I pounded the keys so hard that night that the letter e flew off the part of the machine that hits the paper. Not wanting to waste the night, I went next door to a neighbor who, I knew, had an elaborate workshop in his cellar. He attempted to solder my e back, but when I started to work again, it flew off like a bumblebee. For the rest of the night I inserted each e by hand, and in the morning I took the last dollars from our savings account to buy a new typewriter. Nothing could be allowed to delay the arrival of my greatest triumph.



Figure Ans.19: Image by MS Copilot, Sept. 2025.

—Sloan Wilson, *What Shall We Wear to This Party*, (1976).

The tricky bonus sequence, which is not in the oeis as of April 2024, is based on the lengths of the months. January has 31 days, so 131, and February (often) has 28 days, producing 228. Thus, the missing element is April, 30 days, or 430. Is this neat or nifty?

5.11. We look at the ratio

$$\frac{\frac{n}{n^{3/2}+3}}{\frac{1}{\sqrt{n}}} = \frac{n^{3/2}}{n^{3/2}+3} = \frac{1}{1+3n^{-3/2}} \stackrel{n\to\infty}{\longrightarrow} 1.$$

The p-test implies that the series $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$ diverges, and since the ratio of it and our series is 1, our series also diverges.

5.12. This is trivial. We consider each term of this series a distance. Thus 1 means a distance of 1 unit, 1/2 means a distance of half a unit, and so on. Each term in the series covers half the distance of its predecessor. The total distance covered is 2, because each term covers half of the remaining distance to 2.

6.1. The Ship of Theseus paradox is an antinomy, and its most popular solution is that, looking at the ships as abstract objects, the old and the new ship do occupy the same space at the same time. However, when considering them as material objects, the new ship is not the same as the original one.

6.2. The sorites paradox is similar to the Ship of Theseus paradox. Sorites (or sorós, from the Greek $\sigma\omega\rho\acute{o}\varsigma$, meaning pile, heap, or stack) is a paradox that starts with a heap of sand, from which we remove one grain at a time. Eventually the heap reduces to a single grain of sand, but at what stage is it no longer a heap?

6.3. When in a good mood, a person tends to be optimistic. When in a bad mood, the same person sees things in a different light, forgets the positive aspects and deceives himself about the importance of the negative aspects of the event at hand. At work, a worker may secretly and wrongly believe that they are superior to their co-workers. After an accident, the guilty party often manages to convince themselves that the other party is at fault. A jealous person may suspect his/her spouse of infidelity and may allow such suspicion to overcome him and become a delusion.

6.4. The infinite gift. Imagine a set of n cubes, each having dimensions of $1/\sqrt{n}$. Their total surface area is

$$\sum_{i=1}^{n} \frac{6}{(\sqrt{i})^2} = 6 \sum_{i=1}^{n} \frac{1}{i}.$$

This is the harmonic series which is known to diverge (albeit slowly) and reach any value for a large enough n. Thus, at the limit where n approaches infinity, the total area becomes infinite. On the other hand, the total volume, which is

$$\sum_{i=1}^{n} \frac{1}{(\sqrt{i})^3} = \sum_{i=1}^{n} i^{-3/2},$$

is always finite and in the limit it becomes the Riemann zeta function $\zeta(3/2)$, which converges to $2.61238\ldots$. Hence, it is possible (as well as inexpensive) to fill it up with gifts.

Another example of a mathematical object with is the fractal known as the Koch snowflake. Its area is finite, but its boundary curve is so fragmented, jagged, crinkled, and zigzaggy that its length is infinite.

6.5. They are part of a triplet.

6.6. (By `chatGPT 4o`, May 2024.) The liar paradox arises from a statement made by a person who declares, "I am lying."

Analysis of the Paradox:

1. If the statement is true: If the person is indeed lying, then the statement "I am lying" must be false. This is a contradiction because it means that the statement is both true and false at the same time.

2. If the statement is false: If the person is not lying, then the statement "I am lying" must be true, which again leads to a contradiction because it implies that the person is both telling the truth and lying simultaneously.

Explanation:

This paradox highlights a fundamental problem in logic and language, where a statement refers to itself in a way that creates a loop of self-reference. It shows the complexity and sometimes the limitations of trying to apply strict logical analysis to certain types of statements. The liar paradox is significant in discussions about truth, self-reference, and the foundations of mathematics and logic.

6.7. The best example I could come up with is an unexploded bomb. Such a bomb, if undisturbed, may remain active for many years, but it is reasonable to assume that the chance of it exploding spontaneously drops over time (although it may not exactly halve each year).

6.8. The Mathematica code of Figure Ans.20 produces 0.107143 and 0.0246724, for the areas under the interval and under the new subinterval, respectively, followed by 0.230276 for the probability of choosing a point at random on the curve in the subinterval $[0.75, 0.85]$. This probability is greater than the one for subinterval $[0.55, 0.65]$, because the curve is higher, thus indicating higher probabilities.

```
crv[x_] := x^3 (1. - x^3);
a = Integrate[crv[x], {x, 0, 1.}]
b = Integrate[crv[x], {x, 0.75, 0.85}]
b/a
```

Figure Ans.20: Probability of the Subinterval $[0.75, 0.85]$.

6.9. The number of chords is infinite, but an uncountable infinite. Therefore, the numbers assigned to the chords must be real; there are "not enough" integers for this task. The problem is that it is impossible to assign real numbers to the chords (or any other objects) generated in his way because a real number does not have a successor or a predecessor.

6.10. This is similar to the Bertrand box paradox and should be treated as an example of conditional probability. If we simply assume that someone's child is a boy, then the probability that her other child is also a boy is 1/2, since the other child may be any gender with equal probability. In our particular case, however, we first hear that she has two children, so the possibilities are BB, GG, BG, and GB. We then learn that the boy is her son, which eliminates case GG, and turns the problem into conditional probability. Applying Bayes' rule, the desired probability becomes

$$P(\text{two boys}|\text{at least one boy}) = 1/3.$$

7.1. *The Law of Small Numbers* is an 1898 book (in German) by Ladislaus Bortkiewicz. However, there are the two strong laws of small numbers, observations made by Richard Guy. The first of these [Guy 88], claims that "there aren't enough small numbers to meet the many demands made of them," or, alternatively "You can't tell by looking." The second law [Guy 90], states that "When two numbers look equal, it ain't necessarily so." An example of the first law is Moser's circle sequence, discussed on Page 282.

7.2. In his intriguing book *Death by Black Hole*, reference [Tyson 07], Neil deGrasse Tyson says: "The recoil of Earth from the launch of a single space probe can influence our future in such a way that in about 200 million years, the position of Earth in its orbit around the Sun will be shifted by nearly 60°."

7.3. Such a claim is true in systems that generate random numbers and then remove any number generated. Imagine a card game where the Jack is removed from the game under certain conditions. Once it is removed and the deck is reshuffled, the chances of other cards to appear are increased. However, in a system that produces random numbers without removals, such a claim is wrong. The chance of a number A to be drawn next is constant and does not depend on how often A has been selected earlier. The next roll of a die can produce one of the numbers 1 through 6 regardless of how each has appeared in the past.

Thus, gamblers, which are generally poor decision makers, are sometimes correct (just joking).

> The excitement that a gambler feels when making a bet is equal to the amount he might win times the probability of winning it.
>
> —Blaise Pascal.

7.4. Imagine many successive rolls of a die that rarely result in 4. This simply means that the die is loaded and should be replaced. In a process where probabilities are uniform, each result will appear about the same number of times, and this behavior improves the more numbers are drawn. This behavior is known as the law of large numbers.

> "I think you're begging the question," said Haydock, "and I can see looming ahead one of those terrible exercises in probability where six men have white hats and six men have black hats and you have to work it out by mathematics how likely it is that the hats will get mixed up and in what proportion. If you start thinking about things like that, you would go round the bend. Let me assure you of that!"
>
> —Agatha Christie, *The Mirror Crack'd*, 1962.

7.5. Here are several important examples:

- Radioactive Decay. A radioactive isotope decay exponentially over time. This is a statistical process where the amount of material decreases by a constant fraction over equal time intervals.

- Drug Elimination. The amount of a drug in the body decreases exponentially over time due to metabolism.

- Cooling. The temperature of an object placed in cooler surroundings drops exponentially. The rate of temperature drop is proportional to the temperature difference between the object and its environment.

- Depreciation: Certain objects, most notably cars, depreciate in value over time, and the depreciation may be exponential.

- Populations. The size of an endangered species (or a population affected by disease) declines over time, and the rate of decline can sometimes be modeled with exponential decay.

7.6. Just do it.

7.7. If the die lands on 6, simply roll it again. A similar technique is to use a 10-sided die and have each integer from 1 to 5 written on two sides.

7.8. A sphere can be considered a polyhedron with infinitely many faces and vertices, which is why the chance of a rolled sphere landing on any given point is zero. If we write six numbers equidistant on the surface of a sphere and roll it, chances are the sphere will never stop on any of the numbers.

It is possible to make a spherical die by cutting the sphere in two and creating a cavity in the shape of half an octahedron in each hemisphere. An octahedron has eight faces and six vertices, so each of its halves is a square pyramid. Now place a small, heavy ball inside one cavity and glue the two hemispheres together.

When this sphere is rolled, the heavy ball drops from vertex to vertex and weighs down the sphere such that it eventually stops at one of the six special positions that correspond to the ball being in a vertex.

It would seem that such a sphere is not fair because the cavities cause its inside to be asymmetric and nonuniform. However, with a small, lightweight plastic sphere and a heavy metal ball, the mass of the ball more than compensates for any unfairness caused by the nonuniformity of the sphere.

7.9. In any of the dice of the set, every pair of opposite faces add up to the same number. This is apparent in the standard 6-sided dice used in gambling.

7.10. Here are a number of simple solutions, explained in [harshbarger 24].

- Each player rolls a standard d6 die. If there are ties in the highest or second highest results, they all roll again. This is simple and can accommodate any number of players (simply roll to determine the first among n players, then roll to choose the first among the $n - 1$ remaining players, and so on).

- If there are fewer than 14 players, take the 13 hearts from a deck of cards, shuffle, and have each players pull a card.

- Those who like to work hard can dirty their hands and construct a special n-sided die with the $n!$ permutations of integers or letters. This is simple for $n = 3$, where only a d6 die with the 3! permutations is needed, and may be practical for up to $5! = 120$.

Enough nonsense!

7.11. A good example. Doctor Badcock decides to become a candidate for an important public office, but she has to win an election for this. She appears on television and claims to be ahead of her competitors, but you estimate her chances to be very low, based on the results of public opinion polls and private discussions with your friends and acquaintances.

A bad example. A person buys a lottery ticket every month, even though he knows that his chance of winning are astronomically small. He justifies his bad decision by saying "someone must eventually win, so why not me?"

7.12. It is taken from the middle column of a standard telephone keypad.

7.13. Here is his reasoning: "I see two blacks. Assume that my cross is white, then the other two see one white cross (me) and one black cross. Mr. A saw that we all raised our hands, so he must have said to himself: 'Mr. B, with a black cross on his forehead raised his hand, because he saw a black cross. But since Mr. C's cross is white, B must have seen a black cross on my forehead.' "

"However, Mr. A never said 'stop,' which is why my cross cannot be white. It is black, so I better yell 'stop' before it is too late."

7.14. I don't know the precise answer, but the probability in question must be very low, because it is generally agreed that only 10% of the population (of humans) is left handed, and not much is known about left-handedness in monkeys. The little-known painting *The licentious kitchen maid* (c. 1665) by Pieter van Roestraten depicts such a scene, but this must be because the painter used a camera obscura to project the scene onto his canvas and then trace it before painting the picture.

7.15. He starts by sending messages to a large number of investors. In half the messages he predicts that the stock market will go up in the next six months and in the other half he predicts the opposite. Once the trend of the market becomes known, he deletes from his mailing list the investors who received the wrong message, and sends a similar wave of messages to the remaining half. After a few such steps, his word becomes law to the remaining ones on his list.

7.16. (1) The total number of outcomes of flipping a coin six times is $2^6 = 64$. The number of outcomes with exactly three heads is $\binom{6}{3} = 20$, because this is the number of ways to choose exactly three objects out of a total of six. Thus, the probability in question is $20/64 = 5/16$.

(2) Out of the 64 outcomes of six coin tosses, the following 12 contain three heads in a row:

```
HHHTHT HHHTTH HHHTHH HTHHHT HTTHHH HHTHHH THHHTT THHHTH THTHHH TTHHHT TTTHHH
```

Thus, the probability is $12/64 = 3/16$.

(3) The list above shows that there are only four outcomes with exactly three heads in a row, yielding a probability of $4/64 = 1/16$.

7.17. This is identical to the analysis in the main text, except that x and y change places in equation (7.10). The result is the blue triangle of Figure 7.24b. Thus, the total probability that the three segments could form a triangle is the sum of the areas of the two triangles, which are $1/8$ each.



Figure Ans.21: The Licentious Kitchen Maid (1665).



Figure Ans.22: Image by MS Copilot, Aug. 2025.

7.18. Assume that the length of the rod is 1 unit and points on it have coordinates from 0 to 1. The break point x is equally likely to occur anywhere along the rod, which is why our random variable x, the length of the shorter piece, varies from 0 to 0.5. The break is random, so the distribution of x is uniform over the interval $(0, 0.5)$, which makes the average length of the smaller piece 0.25.

This intuitive result can also be computed formally by integrating the length of the smaller piece, but this should be done carefully, because when the smaller piece starts

at the left end of the rod, its endpoints are at 0 and x, but when it ends at the right end, its endpoints are at $1-x$ and 1. In either case its length is x. The average length is therefore the sum

$$\int_0^{0.5} x\,dx + \int_{0.5}^{1} (1-x)\,dx = \frac{1}{2}x^2\,\Big|_0^{0.5} + \left(x - \frac{x^2}{2}\right)\,\Big|_{0.5}^{1} = 1/8 + 1/8 = 1/4.$$

7.19. Counting the number of distinct visitors to a website. A visitor is identified by their 32-bit IP address, which is then hashed and partitioned as described in the text.

8.1. No. This is because of the choice of axiom.

8.2. Here is GeoGebra code, written by chatGPT.

```
// Parameters
ClearAll[]
depth = 5 // recursion depth
length = 5 // initial trunk length
angle = <value in degrees>  // branching angle

// Recursive function to draw binary tree
BinaryTree[x, y, len, a, d] :=
If(d == 0,
    // base case: do nothing
    "",
    (
        // Compute end point of current branch
        x1 = x + len * cos(a)
        y1 = y + len * sin(a)

        // Draw branch
        Segment[(x, y), (x1, y1)]

        // Recurse on left and right branches
        BinaryTree[x1, y1, len * 0.7, a + angle, d - 1]
        BinaryTree[x1, y1, len * 0.7, a - angle, d - 1]
    )
)
// Call the function from origin
BinaryTree[0, 0, length, 90<degrees>, depth]
// BinaryTree[x, y, len, a, d]
```

8.3. Here is GeoGebra code, written by chatGPT.

```
// Parameters
depth = 4
length = 6

// Clear previous drawings
DeleteAll[]
ZoomIn[(-length, -length), (length, length)]

// Recursive procedure
SierpinskiCarpet[x, y, size, d] :=
```

```
  If(d == 0,
    Polygon[
      (x, y),
      (x + size, y),
      (x + size, y + size),
      (x, y + size)
    ],
    // Subdivide into 9 squares, skip the center
    For[i = 0, i < 3, i++,
      For[j = 0, j < 3, j++,
        If(!(i == 1 && j == 1),
          SierpinskiCarpet[x + i * size / 3, y + j * size / 3, size / 3, d - 1]
        )
      ]
    ]
  )
// Call the function
SierpinskiCarpet[-length / 2, -length / 2, length, depth]
```

8.4. Here is Geogebra code, generated by Microsoft Copilot:

```
// Parameters
depth = 4  // recursion depth
length = 5 // Length of a triangle side

// Initial triangle points
A = (0, 0)
B = (length, 0)
C = (length / 2, length * sqrt(3) / 2)
Polygon(A, B, C)

// Recursive function for a Koch segment
KochSegment(p1, p2, n) =
  If(n == 0,
     Segment(p1, p2),
     {
       d = Distance(p1, p2) / 3,
       angle = Angle(p2 - p1),
       pA = p1,
       pB = p1 + d * UnitVector(p2 - p1),
       pC = pB + d * Rotate(UnitVector(p2 - p1), 60*),
       pD = p1 + 2 * d * UnitVector(p2 - p1),
       KochSegment(pA, pB, n - 1),
       KochSegment(pB, pC, n - 1),
       KochSegment(pC, pD, n - 1),
       KochSegment(pD, p2, n - 1)
     }
  )

// Apply KochSegment to each triangle side
KochSegment(A, B, depth)
KochSegment(B, C, depth)
KochSegment(C, A, depth)
```

* should be “degrees”

Figure Ans.23: Image by chatGPT, Sept. 2025.

This code can be pasted directly into GeoGebra's scripting interface. Alternatively, the CAS view can be employed to define the recursive segments. To see the result in more detail, which may be slower, simply increase the `depth` parameter. An animation can be obtained by repeatedly incrementing `depth` in small steps.

8.5. Both $\tan(x)$ and $\cot(x)$ are good examples, as is also $y = |\sin(x)|$.

8.6. Figure Ans.24a shows a straight line defined by the basic linear interpolation (lerp) relation, Equation (8.6). The canonical parabola $g = x^2$ is also shown. Part (b) of the figure shows how this parabola is inverted, raised, and shifted to the right to become the green curve $h = 1 - (x - 1)^2$. In part (c), the two parabolas are blended linearly to produce the red "ease" curve $(1 - x)g + xh$, which can easily be simplified to $3x^2 - 2x^3$.

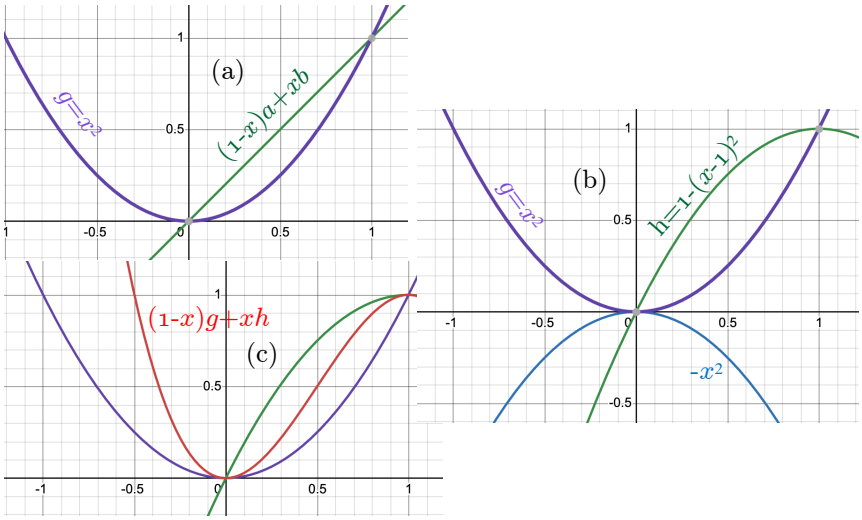


Figure Ans.24: An S-Curve Interpolation of Two Parabolas (Desmos).

8.7. This is easily proved by showing that both dot products $(\mathbf{a}\times\mathbf{b})\bullet\mathbf{a}$ and $(\mathbf{a}\times\mathbf{b})\bullet\mathbf{b}$ are zero. The first one is

$$(\mathbf{a}\times\mathbf{b})\bullet\mathbf{a} = a_1(a_2b_3 - a_3b_2) + a_2(-a_1b_3 + a_3b_1) + a_3(a_1b_2 - a_2b_1) = 0,$$

and similarly for $(\mathbf{a}\times\mathbf{b})\bullet\mathbf{b}$.

8.8. In the special case where $\mathbf{i} = (1, 0, 0)$ and $\mathbf{j} = (0, 1, 0)$, it is easy to verify that the product $\mathbf{i}\times\mathbf{j}$ equals $(0, 0, 1) = \mathbf{k}$. Thus, the triplet $(\mathbf{i}, \mathbf{j}, \mathbf{i}\times\mathbf{j} = \mathbf{k})$ has the handedness of the coordinate system. (It is either right-handed or left-handed, depending on the coordinate system.) In a right-handed coordinate system, the right-hand rule says: If

your thumb points in the direction of $\mathbf{P}$ and your second finger in the direction of $\mathbf{Q}$, then your middle finger will point in the direction of $\mathbf{P}\times\mathbf{Q}$. In a left-handed coordinate system, a similar left-hand rule applies.

8.9. They either point in the same direction or in opposite directions.

8.10. We are looking for a vector $\mathbf{P}(t)$ that's linear in t and that satisfies $\mathbf{P}(0)=\mathbf{P}_1$ and $\mathbf{P}(1)=\mathbf{P}_2$. It is easy to guess that

$$\mathbf{P}(t)=(1-t)\mathbf{P}_1+t\mathbf{P}_2=t(\mathbf{P}_2-\mathbf{P}_1)+\mathbf{P}_1 \qquad \text{(Ans.5)}$$

satisfies both conditions. This equation can also be considered a weighted sum of $\mathbf{P}_1$ and $\mathbf{P}_2$ with the weights $1-t$ and t. It is an important and useful relation often employed in graphics.

8.11. The Mathematica code

```
t14=2^14;
Print["(x*=",(8192-(2*14189.))/t14,",y*=",(14189.+(2*8192))/t14,")"]
Print["(x*=",Cos[60 Degree]-2. Sin[60 Degree],
 ",y*=",Sin[60 Degree]+2. Cos[60 Degree], ")"]
```

calculates the rotated point twice, first using integers and then using Mathematica's built-in sine and cosine functions. The results are identical: ($x^*=-1.23206, y^*=1.86603$).

For an 80° rotation, the code

```
t14=2^14;
Print["(x*=",(2845.-(2*16135.))/t14,",y*=",(16135.+(2*2845.))/t14,
  ")"]
Print["(x*=",Cos[80 Degree]-2. Sin[80 Degree],
 ",y*=",Sin[80 Degree]+2. Cos[80 Degree], ")"]
```

produces ($x^*=-1.79596, y^*=1.33209$) and ($x^*=-1.79597, y^*=1.3321$) (a slightly different result).

8.12. From the definition of θ_i, we know that the ratio $\tan\theta_{i+1}/\tan\theta_i$ is 1/2. Small angles satisfy $\tan\theta\approx\theta$, so we conclude that the ratio θ_{i+1}/θ_i equals approximately 1/2, except for the first few θ_i's. This can also be confirmed by manually checking the ratios from Table 8.37. Given an infinite sequence of numbers t, $t/2$, $t/4$,..., $t/2^i$, we can express every number from 0 (which is obtained by subtracting all the numbers in the sequence from the first one) to $2t$ (which is obtained by adding all the numbers in the sequence). Our sequence of θ_i is finite and the ratio of consecutive elements isn't always precisely 1/2, but [Walther 71] proves that every number in the range $[0, 90°)$ can be reached, up to a certain precision, by adding and subtracting a number of consecutive θ_i's.

8.13. The method proposed here is based on the fact that the magnitude of the rotated vector (x^*, y^*) should be identical to that of the original vector (x, y). This can be achieved by first normalizing (x^*, y^*) and then multiplying it by the magnitude of (x, y),

$$(x^*, y^*) \leftarrow (x^*, y^*)\frac{\sqrt{x^2+y^2}}{\sqrt{x^{*2}+y^{*2}}} = (x^*, y^*)\sqrt{\frac{x^2+y^2}{x^{*2}+y^{*2}}},$$

a calculation that requires four exponentiations, one division, one multiplication, and one square root.

8.14. The traditional way of calculating a sine function is by its power series

$$\sin(\theta) = \frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots,$$

and similarly for cosine. These series, however, converge very slowly, requiring many multiplications and divisions. If a graphics application needs just rotations, the method of Section 8.8.7 may be simpler and faster than CORDIC. The advantage of CORDIC is that it can be adapted to the calculation of many different functions. A general software package that is concerned not just with rotations may benefit from the application of CORDIC.

8.15. We can write $\mathbf{P}_1 = \mathbf{P}_0 + \alpha(\mathbf{P}_3 - \mathbf{P}_0)$ and similarly $\mathbf{P}_2 = \mathbf{P}_0 + \beta(\mathbf{P}_3 - \mathbf{P}_0)$. It is now obvious that n collinear points can be represented by two points and $n-2$ real parameters.

8.16. A quadratic polynomial is fully defined by three coefficients $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. They can be two- or three-dimensional points, but any three points are always on the same plane.

8.17. We can gain a deeper insight into the shape of the n-degree polynomial $P(x) = \sum_{i=0}^{n} A_i x^i$ by writing the equation $P(x) = 0$. This is an nth-degree equation in the unknown x and it therefore has n solutions (some may be identical or complex). Each solution is an x value for which the polynomial becomes zero. As x is varied, the polynomial crosses the x axis n times, so it oscillates between positive and negative values, resulting in wiggling.

8.18. This is the case $n = 1$. The general form of the LP is, therefore, $y = \sum_{i=0}^{1} y_i L_i^1$. The weight functions are easy to calculate:

$$L_0^1 = \frac{x - x_1}{x_0 - x_1}, \qquad L_1^1 = \frac{x - x_0}{x_1 - x_0},$$

and the curve is therefore

$$\begin{aligned} y &= y_0 L_0^1 + y_1 L_1^1 = y_0\frac{x - x_1}{x_0 - x_1} + y_1\frac{x - x_0}{x_1 - x_0} \\ &= x\frac{y_0 - y_1}{x_0 - x_1} + \frac{y_1 x_0 - y_0 x_1}{x_0 - x_1} = ax + b, \end{aligned}$$

which is a straight line.

8.19. Since the three points are approximately equally spaced, it makes sense to use knot values $t_0 = 0$, $t_1 = 1/2$, and $t_2 = 1$. The first step is to calculate the three basis functions $L_i^2(t)$:

$$\begin{aligned}
L_0^2 &= \frac{\Pi_{j\neq 0}^2(t-t_j)}{\Pi_{j\neq 0}^2(t_i-t_j)} = \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} = 2(t-1/2)(t-1),\\
L_1^2 &= \frac{\Pi_{j\neq 1}^2(t-t_j)}{\Pi_{j\neq 1}^2(t_i-t_j)} = \frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)} = -4t(t-1),\\
L_2^2 &= \frac{\Pi_{j\neq 2}^2(t-t_j)}{\Pi_{j\neq 2}^2(t_i-t_j)} = \frac{(t-t_0)(t-t_1)}{(t_2-t_0)(t_2-t_1)} = 2t(t-1/2).
\end{aligned}$$

The LP is now easy to calculate:

$$\begin{aligned}
\mathbf{P}(t) &= (0,0)2(t-1/2)(t-1) - (0,1)4t(t-1) + (1,1)2t(t-1/2)\\
&= (2t^2-t, -2t^2+3t). \qquad \text{(Ans.6)}
\end{aligned}$$

This is a quadratic (degree-2) parametric polynomial and a simple test verifies that it passes through the three given points.

8.20. The first step is to calculate the basis functions

$$N_0(t) = 1, \quad N_1(t) = t - t_0 = t, \quad N_2(t) = (t-t_0)(t-t_1) = t(t-1/2).$$

The next step is to compute the three coefficients

$$\begin{aligned}
\mathbf{A}_0 &= \mathbf{P}_0 = (0,0),\\
\mathbf{A}_1 &= \frac{\mathbf{P}_1 - \mathbf{P}_0}{t_1 - t_0} = \frac{(0,1)-(0,0)}{1/2} = (0,2),\\
\mathbf{A}_2 &= \frac{\dfrac{(1,1)-(0,1)}{1-1/2} - \dfrac{(0,1)-(0,0)}{1/2-0}}{1-0} = (2,-2).
\end{aligned}$$

The polynomial can now be calculated:

$$\mathbf{P}(t) = 1\times(0,0) + t(0,2) + t(t-1/2)(2,-2) = (2t^2-t, -2t^2+3t).$$

It is, of course, identical to the LP calculated in Exercise 8.19.

8.21. The curve is given by $\mathbf{P}(t) = (2t^2-t, -2t^2+3t)$, so its derivative is $\mathbf{P}^t(t) = (4t-1, -4t+3)$. The three tangent vectors are $\mathbf{P}^t(t_0 = 0) = (-1,3)$, $\mathbf{P}^t(t_1 = 1/2) = (1,1)$, and $\mathbf{P}^t(t_2 = 1) = (3,-1)$. The direction of tangent vector $(-1,3)$ is described by saying "for every three steps in the y direction, the curve moves one step in the negative x direction."

The slopes are calculated by dividing the y coordinate of a tangent vector by its x coordinate. The slopes at the three points are therefore $-3/1$, $1/1$, and $-1/3$. They correspond to angles of 288.44°, 45°, and −18.43°, respectively.

8.22. When the user specifies four points, the curve should pass through the original points. After a point is moved, the curve will no longer pass through the original point. When only the two endpoints are specified, the user is normally willing to consider different curves that pass through them, with different start and end directions.

8.23. Take one of these vectors, say, $(2, 1, 0.6)$, and divide it by its magnitude. The result is

$$\frac{(2,1,0.6)}{\sqrt{2^2+1^2+0.6^2}} \approx \frac{(2,1,0.6)}{2.93} = (0.7272, 0.3636, 0.2045).$$

The new vector points in the same direction but its magnitude is 1. Its components therefore satisfy

$$\sqrt{0.7272^2+0.3636^2+0.2045^2} = 1, \text{ or } 0.7272^2+0.3636^2+0.2045^2 = 1, \qquad \text{(Ans.7)}$$

so they are dependent. Any of them can be calculated from the other two with Equation (Ans.7).

8.24. Substituting $t = 0.5$ in Equation (8.29) yields

$$\begin{aligned}\mathbf{P}(0.5) &= (2\mathbf{P}_1 - 2\mathbf{P}_2 + \mathbf{P}_1^t + \mathbf{P}_2^t)/8 + (-3\mathbf{P}_1 + 3\mathbf{P}_2 - 2\mathbf{P}_1^t - \mathbf{P}_2^t)/4 + \mathbf{P}_1^t/2 + \mathbf{P}_1 \\ &= \frac{1}{2}(\mathbf{P}_1 + \mathbf{P}_2) + \frac{1}{8}(\mathbf{P}_1^t - \mathbf{P}_2^t). \qquad \text{(Ans.8)}\end{aligned}$$

The first part of this expression is the midpoint of the segment $\mathbf{P}_1 \to \mathbf{P}_2$ and the second part is the difference of the two tangents, divided by 8. Figure Ans.25 illustrates how adding $(\mathbf{P}_1^t - \mathbf{P}_2^t)/8$ to the midpoint of $\mathbf{P}_1 \to \mathbf{P}_2$ brings us to the midpoint of the curve.

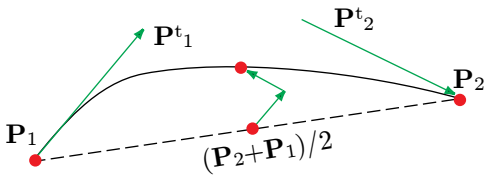


Figure Ans.25: The Midpoint $\mathbf{P}(0.5)$ of a Hermite Segment.

8.25. The Hermite segment is a cubic polynomial in t, so its third derivative is constant. It is easy to see, from Equation (8.31), that the third derivatives of the Hermite blending functions $F_i(t)$ are

$$F_1^{ttt}(t) = 12, \quad F_2^{ttt}(t) = -12, \quad F_3^{ttt}(t) = 6, \quad F_4^{ttt}(t) = 6.$$

The third derivative of the segment is therefore

$$\begin{aligned}\mathbf{P}^{ttt}(t) &= (12\mathbf{P}_1 - 12\mathbf{P}_2 + 6\mathbf{P}_1^t + 6\mathbf{P}_2^t) \\ &= (t^3, t^2, t, 1)\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 12 & -12 & 6 & 6 \end{bmatrix}\begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_1^t \\ \mathbf{P}_2^t \end{bmatrix} \\ &= \mathbf{T}(t)\mathbf{H}_{ttt}\mathbf{B} = \mathbf{H}_{ttt}\mathbf{B}.\end{aligned}$$

$\mathbf{P}^{ttt}(t)$ is independent of t, because the top three rows of $\mathbf{H}_{ttt}$ are zero. This derivative is the constant vector $12(\mathbf{P}_1 - \mathbf{P}_2) + 6(\mathbf{P}_1^t + \mathbf{P}_2^t)$.

Here are the Hermite matrix and its derivatives side by side. Use your experience to explain how each is derived from its predecessor.

$$\mathbf{H} = \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{H}_t = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 6 & -6 & 3 & 3 \\ -6 & 6 & -4 & -2 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\mathbf{H}_{tt} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 12 & -12 & 6 & 6 \\ -6 & 6 & -4 & -2 \end{pmatrix}, \quad \mathbf{H}_{ttt} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 12 & -12 & 6 & 6 \end{pmatrix}.$$

8.26. It's trivial to show that $\mathbf{P}(0) = (-1,0)0^3 + (1,-1)0^2 + (1,1)0 = (0,0)$ and $\mathbf{P}(1) = (-1,0)1^3 + (1,-1)1^2 + (1,1)1 = (1,0)$. The tangent vector of $\mathbf{P}(t)$ is

$$\frac{d\,\mathbf{P}(t)}{d\,t} = 3(-1,0)t^2 + 2(1,-1)t + (1,1),$$

so the two extreme tangent vectors are

$$\begin{aligned}\frac{d\,\mathbf{P}(0)}{d\,t} &= 3(-1,0)0^2 + 2(1,-1)0 + (1,1) = (1,1), \\ \frac{d\,\mathbf{P}(1)}{d\,t} &= 3(-1,0)1^2 + 2(1,-1) + (1,1) = (0,-1),\end{aligned}$$

as should be.

8.27. Similar to the previous example, we get

$$\begin{aligned}\mathbf{P}(t) &= (t^3, t^2, t, 1)\mathbf{H}\left((0,0),(1,0),(2,2),(0,-1)\right)^T \\ &= (0,1)t^3 - (1,3)t^2 + (2,2)t.\end{aligned}$$

It's a different polynomial and it has a different shape; yet a simple check shows that it passes through the same endpoints and has the same start and end directions.

8.28. Equation (8.32) becomes

$$\mathbf{P}(t)=(t^3,t^2,t,1)\begin{bmatrix}2&-2&1&1\\-3&3&-2&-1\\0&0&1&0\\1&0&0&0\end{bmatrix}\begin{bmatrix}\mathbf{P}_1\\\mathbf{P}_2\\(0,0)\\(0,0)\end{bmatrix}=(3t^2-2t^3)(\mathbf{P}_2-\mathbf{P}_1)+\mathbf{P}_1. \quad \text{(Ans.9)}$$

To find the type of the curve, we substitute $j=3t^2-2t^3$ (note that $t=0\Rightarrow j=0$ and $t=1\Rightarrow j=1$). This results in the familiar expression $\mathbf{P}(t)=j(\mathbf{P}_2-\mathbf{P}_1)+\mathbf{P}_1=(1-j)\mathbf{P}_1+j\mathbf{P}_2$. The curve is therefore the straight segment from $\mathbf{P}_1$ to $\mathbf{P}_2$. The (important) conclusion is: If the initial and final directions of the Hermite segment are not specified, the curve will "choose" the shortest path from $\mathbf{P}_1$ to $\mathbf{P}_2$.

8.29. For case 1, we use the notation $\mathbf{P}^t(0)=\mathbf{P}_1^t$, $\mathbf{P}^t(1/2)=\mathbf{P}_2^t$, and $\mathbf{P}^t(1)=\mathbf{P}_3^t$. From $\mathbf{P}(t)=\mathbf{a}t^3+\mathbf{b}t^2+\mathbf{c}t+\mathbf{d}$, we get $\mathbf{P}^t(t)=3\mathbf{a}t^2+2\mathbf{b}t+\mathbf{c}$, resulting in the three equations

$$\begin{aligned}3\mathbf{a}\cdot 0^2+2\mathbf{b}\cdot 0+\mathbf{c}&=\mathbf{P}_1^t,\\3\mathbf{a}\cdot(1/2)^2+2\mathbf{b}\cdot(1/2)+\mathbf{c}&=\mathbf{P}_2^t,\\3\mathbf{a}\cdot 1^2+2\mathbf{b}\cdot 1+\mathbf{c}&=\mathbf{P}_3^t,\end{aligned}$$

where the unknowns are $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$ (notice that $\mathbf{d}$ does not participate in our equations). It is clear that $\mathbf{c}=\mathbf{P}_1^t$. The other two unknowns are solved by the simple Mathematica code `Solve[{3a/4+2b/2+p1==p2, 3a+2b+p1==p3}, {a,b}]`, which yields $\mathbf{a}=\frac{2}{3}(\mathbf{P}_1^t-2\mathbf{P}_2^t+\mathbf{P}_3^t)$ and $\mathbf{b}=\frac{1}{2}(-3\mathbf{P}_1^t+4\mathbf{P}_2^t-\mathbf{P}_3^t)$. Thus, the curve is given by

$$\begin{aligned}\mathbf{P}(t)&=\mathbf{a}t^3+\mathbf{b}t^2+\mathbf{c}t+\mathbf{d}\\&=\frac{2}{3}(\mathbf{P}_1^t-2\mathbf{P}_2^t+\mathbf{P}_3^t)t^3+\frac{1}{2}(-3\mathbf{P}_1^t+4\mathbf{P}_2^t-\mathbf{P}_3^t)t^2+\mathbf{P}_1^t t+\mathbf{d},\end{aligned}$$

which shows that the three given tangents fully determine the shape of the curve but not its position in space. The latter requires the value of $\mathbf{d}$.

For case 2, we denote $\mathbf{P}(1/3)=\mathbf{P}_1$, $\mathbf{P}(2/3)=\mathbf{P}_2$, $\mathbf{P}^t(0)=\mathbf{P}_1^t$, and $\mathbf{P}^t(1)=\mathbf{P}_2^t$. This results in the four equations

$$\begin{aligned}\mathbf{a}(1/3)^3+\mathbf{b}(1/3)^2+\mathbf{c}(1/3)+\mathbf{d}&=\mathbf{P}_1,\\\mathbf{a}(2/3)^3+\mathbf{b}(2/3)^2+\mathbf{c}(2/3)+\mathbf{d}&=\mathbf{P}_2,\\3\mathbf{a}\cdot 0^2+2\mathbf{b}\cdot 0+\mathbf{c}&=\mathbf{P}_1^t,\\3\mathbf{a}\cdot 1^2+2\mathbf{b}\cdot 1+\mathbf{c}&=\mathbf{P}_2^t,\end{aligned}$$

where the unknowns are again $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$. It is again clear that $\mathbf{c}=\mathbf{P}_1^t$ and the other three unknowns are easily solved by the code

```
Solve[{a (1/3)^3+b (1/3)^2+p1t (1/3)+d==p1,
a (2/3)^3+b (2/3)^2+p1t (1/3)+d==p2, 3a+2b+p1t==p2t}, {a,b,d}],
```

which yields the solutions

$$\begin{aligned}\mathbf{a} &= -\frac{9}{13}(-6\mathbf{P}_1 + \mathbf{P}_1^t + 6\mathbf{P}_2 - \mathbf{P}_2^t),\\ \mathbf{b} &= \frac{1}{13}(-81\mathbf{P}_1 + 7\mathbf{P}_1^t + 81\mathbf{P}_2 - 7\mathbf{P}_2^t),\\ \mathbf{d} &= \frac{1}{117}(180\mathbf{P}_1 - 43\mathbf{P}_1^t - 63\mathbf{P}_2 + 4\mathbf{P}_2^t).\end{aligned}$$

Thus, the PC segment is

$$\begin{aligned}\mathbf{P}(t) &= \mathbf{a}t^3 + \mathbf{b}t^2 + \mathbf{c}t + \mathbf{d}\\ &= -\frac{9}{13}(-6\mathbf{P}_1 + \mathbf{P}_1^t + 6\mathbf{P}_2 - \mathbf{P}_2^t)t^3 + \frac{1}{13}(-81\mathbf{P}_1 + 7\mathbf{P}_1^t + 81\mathbf{P}_2 - 7\mathbf{P}_2^t)t^2\\ &\quad + \mathbf{P}_1^t \cdot t + \frac{1}{117}(180\mathbf{P}_1 - 43\mathbf{P}_1^t - 63\mathbf{P}_2 + 4\mathbf{P}_2^t).\end{aligned}$$

Case 3 is similar to case 2 and is not shown here.

8.30. By using the *same* symbol, $\mathbf{P}_{k+1}^t$, for the end tangent of $\mathbf{P}_k(t)$ and the start tangent of $\mathbf{P}_{k+1}(t)$.

8.31. The code and curve of Figure Ans.26 shows that the first segment is very tight. It is does not proceed straight to the right because it has to reach point $(1, 0)$ at an angle, but it is the shortest possible. It is also initially slow. A simple test shows that at $t = 0.5$, the curve hasn't reached the midpoint between $\mathbf{P}_1$ and $\mathbf{P}_2$.

8.32. When $T > 1$, s becomes negative, causing the two direction vectors to reverse directions. This changes the shape of the curve completely. However, large negative values of s still produce a loose curve.

8.33. The tangent vector of the curve is easily calculated from Equation (8.49):

$$\mathbf{P}_1^t(t) = (-1.5t^2 + 2t - 0.5)\mathbf{P}_1 + (4.5t^2 - 5t)\mathbf{P}_2 + (-4.5t^2 + 4t + 0.5)\mathbf{P}_3 + (1.5t^2 - 1)\mathbf{P}_4.$$

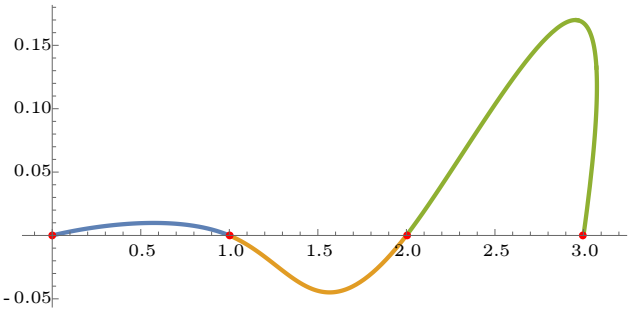
At the end ($t = 1$), the tangent is $-0.5\mathbf{P}_2 + 0.5\mathbf{P}_4$.

The tangent vector of the next segment has the same coefficients, so its form is

$$\mathbf{P}_2^t(t) = (-1.5t^2 + 2t - 0.5)\mathbf{P}_2 + (4.5t^2 - 5t)\mathbf{P}_3 + (-4.5t^2 + 4t + 0.5)\mathbf{P}_4 + (1.5t^2 - 1)\mathbf{P}_5.$$

At the start ($t = 0$), this tangent also has the value $-0.5\mathbf{P}_2 + 0.5\mathbf{P}_4$, so the two tangents are equal at the connection points between curve segments.

8.34. Figure Ans.27 lists the points and the code for this computation. Notice how the sharp corner at the top-center of the heart is obtained by the particular placement of points 3 through 6 and how parameter `ppr` determines the width of the heart.




```
p1 = {0, 0}; p2 = {1, 0}; p3 = {2, 0}; p4 = {3, 0};
mt = {{1, 4, 1, 0}, {0, 1, 4, 1}};
c = {{{{0, 0}}}, {{{a11, a12}}}, {{{a21, a22}}}, {{{-1, -1}}}};
r = {{3 ({p3} - {p1})}, {3 ({p4} - {p2})}};
Solve[mt . c == r, {a11, a12, a21, a22}]
pt1 = {0, 0}; pt2 = {17/15, -1/15};
pt3 = {22/15, 4/15}; pt4 = {-1, -1};
tr = {t^3, t^2, t, 1};
m4 = {{2, -2, 1, 1}, {-3, 3, -2, -1}, {0, 0, 1, 0}, {1, 0, 0, 0}};
po1[t_] := tr . m4 . {{{p1}}, {{p2}}, {{pt1}}, {{pt2}}};
po2[t_] := tr . m4 . {{{p2}}, {{p3}}, {{pt2}}, {{pt3}}};
po3[t_] := tr . m4 . {{{p3}}, {{p4}}, {{pt3}}, {{pt4}}};
g1 = ParametricPlot[{po1[t], po2[t], po3[t]}, {t, 0, 1}];
g2 = ListPlot[{p1, p2, p3, p4}, PlotStyle -> {PointSize[Medium], Red}];
Show[g1, g2, AspectRatio -> .5]

{{a11->17/15,a12->-(1/15),a21->22/15,a22->4/15}}
```

Figure Ans.26: A Tight Cubic Spline Example.

8.35. We simply calculate the quadratic Bézier curve for the three points. As a quadratic parametric polynomial it is a parabola (see second paragraph of Section 8.11). Since this is a Bézier curve, its extreme tangents point in the desired directions:

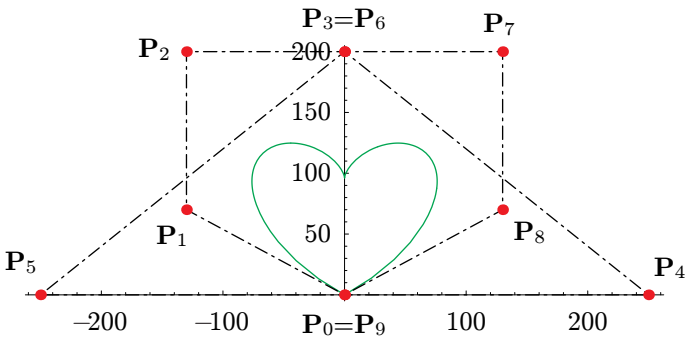
$$\mathbf{P}(t) = \mathbf{P}_1(1-t)^2 + 2\mathbf{P}_2(1-t)t + \mathbf{P}_3t^2 = (\mathbf{P}_1 - 2\mathbf{P}_2 + \mathbf{P}_3)t^2 + 2(\mathbf{P}_2 - \mathbf{P}_1)t + \mathbf{P}_1.$$

8.36. A simple procedure is to compute

$$\mathbf{P}_0 = \mathbf{P}(0) = (1,0), \quad \mathbf{P}_1 = \mathbf{P}(1/3) = (13/9, 1/27),$$
$$\mathbf{P}_2 = \mathbf{P}(2/3) = (19/9, 8/27), \quad \mathbf{P}_3 = \mathbf{P}(1) = (3,1).$$

8.37. The substitution is $u = 2t - 1$, from which we get $t = (1+u)/2$ and $1 - t = (1-u)/2$. The curve of Equation (8.58) can now be written

$$\mathbf{P}(t) = \frac{1}{8}(1-u)^3\mathbf{P}_0 + \frac{1}{4}(1+u)(1-u)^2\mathbf{P}_1 + 2\left(\frac{1+u}{2}\right)^2\left(\frac{1-u}{2}\right)\mathbf{P}_2 + \frac{1}{8}(1+u)^3\mathbf{P}_3$$




```
(*Heart-shaped Bezier curve*)n=9;ppr=130;
pnts={{0,0},{-ppr,70},{-ppr,200},{0,200},{250,0},{-250,0},
 {0,200},{ppr,200},{ppr,70},{0,0}};
pwr[x_,y_]:=If[x==0&&y==0,1,x^y];
bern[n_,i_,t_]:=Binomial[n,i]pwr[t,i]pwr[1-t,n-i]
bzCurve[t_]:=Sum[pnts[[i+1]]bern[n,i,t],{i,0,n}]
g1=ListPlot[pnts,PlotStyle->{Red,AbsolutePointSize[6]}];
g2=ParametricPlot[bzCurve[t],{t,0,1}];
g3=Graphics[{AbsoluteDashing[{1,2,5,2}],Line[pnts]}];
Show[g1,g2,g3,PlotRange->All]
```

Figure Ans.27: A Heart-Shaped Bézier Curve.

$$= \frac{1}{8}(u^3, u^2, u, 1) \begin{pmatrix} -1 & 2 & -2 & 1 \\ 3 & -2 & -2 & 3 \\ -3 & -2 & 2 & 3 \\ 1 & 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{P}_0 \\ \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_3 \end{pmatrix}.$$

The only difference is the basis matrix.

8.38. Direct calculation of $B_{4,i}(t)$ for $0 \le i \le 4$ yields the five functions

$$B_{4,0} = (1-t)^4,\ B_{4,1} = 4t(1-t)^3,\ B_{4,2} = 6t^2(1-t)^2,\ B_{4,3} = 4t^3(1-t), \text{ and } B_{4,4} = t^4.$$

8.39. Three collinear points are dependent, which means that any of the three can be expressed as a linear combination (a weighted sum) of the other two, with barycentric weights. We therefore assume that $\mathbf{P}_1 = (1-\alpha)\mathbf{P}_0 + \alpha\mathbf{P}_2$ for some real α. The general Bézier curve for three points,

$$\mathbf{P}(t) = \mathbf{P}_0(1-t)^2 + \mathbf{P}_1 2t(1-t) + \mathbf{P}_2 t^2,$$

now becomes

$$\mathbf{P}(t) = \mathbf{P}_0(1-t)^2 + [(1-\alpha)\mathbf{P}_0 + \alpha\mathbf{P}_2]2t(1-t) + \mathbf{P}_2 t^2,$$

which is easily simplified to

$$\begin{aligned}\mathbf{P}(t) &= \mathbf{P}_0 + 2\alpha(\mathbf{P}_2 - \mathbf{P}_0)t + (1-2\alpha)(\mathbf{P}_2 - \mathbf{P}_0)t^2 \\ &= \mathbf{P}_0 + (\mathbf{P}_2 - \mathbf{P}_0)[2\alpha t + (1-2\alpha)t^2] \\ &= \mathbf{P}_0 + (\mathbf{P}_2 - \mathbf{P}_0)T. \end{aligned} \quad \text{(Ans.10)}$$

This is linear in T and therefore represents a straight line.

This case does not contradict the fact that the Bézier curve does not pass through the intermediate points. We have considered three *collinear* points, which really are only two points. The Bézier curve for two points is a straight line. Note that even with four collinear points, only two are really independent.

We continue this discussion by examining two cases. The first is the special case of uniformly-spaced collinear points and the second is the case of three collinear points $\mathbf{P}_0$, $\mathbf{P}_1$, and $\mathbf{P}_2$ where $\mathbf{P}_1$ is not between $\mathbf{P}_0$ and $\mathbf{P}_2$ but is one of the endpoints.

Case 1. Consider the case of $n+1$ points that are equally spaced along the straight segment from $\mathbf{P}_0$ to $\mathbf{P}_n$. We show that the Bézier curve for these points is the straight segment from $\mathbf{P}_0$ to $\mathbf{P}_n$. We start with two auxiliary relations;

1. Point $\mathbf{P}_k$ (for $k = 0, 1, \ldots, n$) can be expressed in this case as the blend $(1-k/n)\mathbf{P}_0 + (k/n)\mathbf{P}_n$.

2. It can be proved by induction that $\sum_{i=0}^{n} iB_{n,i}(t) = nt$.

Based on these relations, the Bézier curve for uniformly-spaced collinear points is

$$\begin{aligned}\mathbf{P}(t) &= \sum_{i=0}^{n} B_{n,i}(t)\mathbf{P}_i = \sum_{i=0}^{n} B_{n,i}(t)\big[(1-i/n)\mathbf{P}_0 + (i/n)\mathbf{P}_n\big] \\ &= \mathbf{P}_0 \sum B_{n,i}(t) - \frac{\mathbf{P}_0}{n}\sum iB_{n,i}(t) + \frac{\mathbf{P}_n}{n}\sum iB_{n,i}(t) \\ &= \mathbf{P}_0 - t\mathbf{P}_0 + t\mathbf{P}_n = (1-t)\mathbf{P}_0 + t\mathbf{P}_n.\end{aligned}$$

Case 2. $\mathbf{P}_1$ is not located between $\mathbf{P}_0$ and $\mathbf{P}_2$ but is one of the endpoints. The two cases $\alpha = 0$ and $\alpha = 1$ imply that point $\mathbf{P}_1$ is identical to $\mathbf{P}_0$ or $\mathbf{P}_2$, respectively. The case $\alpha = 0.5$ means that $\mathbf{P}_1$ is midway between $\mathbf{P}_0$ and $\mathbf{P}_2$. The cases $\alpha < 0$ and $\alpha > 1$ are special. The former means that $\mathbf{P}_1$ "precedes" $\mathbf{P}_0$. The latter means that $\mathbf{P}_1$ "follows" $\mathbf{P}_2$. In these cases, the curve is no longer a straight line but goes from $\mathbf{P}_0$ toward $\mathbf{P}_1$, reverses direction without reaching $\mathbf{P}_1$, and continues to $\mathbf{P}_2$. The point where it reverses direction becomes a cusp (a sharp corner), where the curve has an indefinite tangent vector (Figure Ans.28).

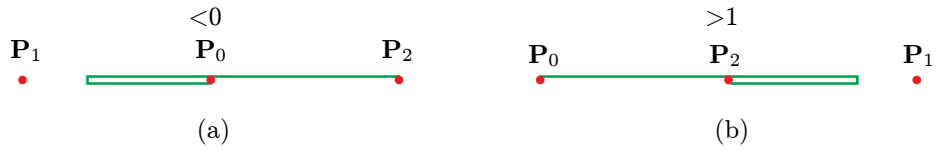


Figure Ans.28: Bézier Straight Segments.

Analysis. We first show that in these cases the curve does not go through point $\mathbf{P}_1$. Equation (Ans.10) can be written

$$\mathbf{P}(t) = \mathbf{P}_0\left(1 - 2\alpha t - t^2 + 2\alpha t^2\right) + \mathbf{P}_2\left(2\alpha t + t^2 - 2\alpha t^2\right).$$

Let's see for what value of t the curve passes through point $\mathbf{P}_1 = (1-\alpha)\mathbf{P}_0 + \alpha\mathbf{P}_2$. The conditions are

$$1 - 2\alpha t - t^2 + 2\alpha t^2 = 1 - \alpha \quad \text{and} \quad 2\alpha t + t^2 - 2\alpha t^2 = \alpha.$$

These conditions yield the following quadratic equations for t:

$$\alpha - 2\alpha t + (2\alpha - 1)t^2 = 0 \quad \text{and} \quad -\alpha + 2\alpha t - (2\alpha - 1)t^2 = 0.$$

These equations are identical and their solutions are

$$t = \frac{\alpha \pm \sqrt{\alpha(\alpha-1)}}{\alpha} \quad \text{and} \quad t = \frac{-\alpha \pm \sqrt{\alpha(1-\alpha)}}{-\alpha}.$$

The first solution has no real values for negative α and the second one has no real values for $\alpha > 1$. For these values of α, the curve does not pass through control point $\mathbf{P}_1$.

We now calculate the value of t for which the curve has a cusp (a sharp corner). The tangent vector of the curve is

$$\mathbf{P}^t(t) = \mathbf{P}_0\left(-2\alpha - 2t + 4\alpha t\right) + \mathbf{P}_2\left(2\alpha + 2t - 4\alpha t\right) = (2\alpha + 2t - 4\alpha t)(\mathbf{P}_2 - \mathbf{P}_0).$$

The condition for an indefinite tangent vector is therefore $2\alpha + 2t - 4\alpha t = 0$, which happens for $t = \alpha/(2\alpha - 1)$.

The following three special cases are particularly interesting:

1. $\alpha \ll 0$. This is the case where $\mathbf{P}_1$ is far away from both $\mathbf{P}_0$ and $\mathbf{P}_2$. The limit of $\alpha/(2\alpha - 1)$ in this case is $1/2$, which means that the curve changes direction at its midpoint.

2. $\alpha = -1$. In this case point $\mathbf{P}_0$ is exactly between $\mathbf{P}_1$ and $\mathbf{P}_2$. The value of $\alpha/(2\alpha - 1)$ in this case is $1/3$ (Figure Ans.28a illustrates why this makes sense).

3. $\alpha \gg 1$. Here, $\mathbf{P}_1$ is again far from both $\mathbf{P}_0$ and $\mathbf{P}_2$, but in the other direction (Figure Ans.28b). The limit of $\alpha/(2\alpha - 1)$ in this case is, again, $1/2$.

(End of long answer.)

8.40. Figure Ans.29 illustrates the construction. First, the blue lines a and b are constructed, followed by the two lines labeled c. This is followed by the ten green lines, five of which are equally spaced on the left-hand side of b and the other five equally spaced on the right-hand side of b. The last step is to construct the eight red vertical line segments.

> We shall therefore borrow all our rules for the finishing of our proportions, from the musicians, who are the greatest masters of this sort of numbers, and from those things wherein nature shows herself most excellent and compleat.
>
> —Leon Battista Alberti.

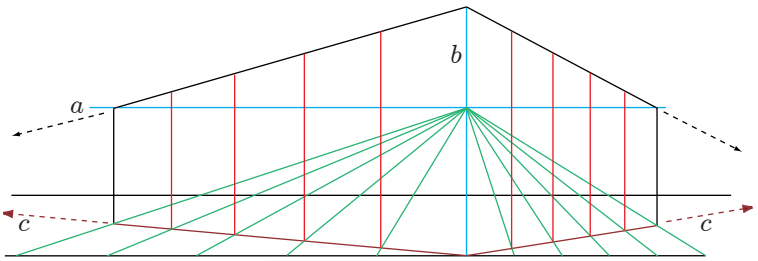


Figure Ans.29: Two-Point Perspective with Equally-Spaced Lines.

8.41. Yes, by viewing it through a telescope. A telescope is a device that "telescopes" a scene and brings objects closer to the observer rather than magnifying them, but it does not affect the perspective.

8.42. In the standard position, the line of sight of the viewer is the z axis. In order for a line segment to be perpendicular to this direction, all its points must have the same z coordinate (i.e., the segment must be contained in a plane parallel to the xy plane). We therefore select two endpoints with $z = 1$ and two other endpoints with $z = 3$. The first two points are selected, somewhat arbitrarily, as $\mathbf{P}_1 = (2, 3, 1)$ and $\mathbf{P}_2 = (3, -1, 1)$. The third point is chosen as $\mathbf{P}_3 = (0, 2, 3)$ and the last point is determined from $\mathbf{P}_4 = \mathbf{P}_2 - \mathbf{P}_1 + \mathbf{P}_3 = (1, -2, 3)$. The four points are now projected to $\mathbf{P}_1^* = (1, 3/2)$, $\mathbf{P}_2^* = (3/2, -1/2)$, $\mathbf{P}_3^* = (0, 1/2)$, and $\mathbf{P}_4^* = (1/4, -1/2)$.

We show that the two straight segments defined by the four projected points are parallel, by computing the differences $\mathbf{v}_1 = \mathbf{P}_2^* - \mathbf{P}_1^* = (1/2, -2)$ and $\mathbf{v}_2 = \mathbf{P}_4^* - \mathbf{P}_3^* = (1/4, -1)$. The difference of two points is a vector, and the two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ point in the same direction.

8.43. We are looking for a t value for which $\mathbf{P}^*(t) = (0, 1/4)$. This can be written as the vector equation

$$(1-t)^2(-1/2, 0) + 2t(1-t)(0, 1/3) + t^2(1/4, 1/4) = (0, 1/4),$$

or as the two separate scalar equations $(1-t)^2(-1/2) + 2t(1-t)(0) + t^2(1/4) = 0$ and $(1-t)^2(0) + 2t(1-t)(1/3) + t^2(1/4) = (1/4)$. The first equation yields the solutions $t \approx 0.5858$ and $t \approx 3.414$, while the second equation has the solutions $t = 0$ and $t = 1.6$. The two equations are therefore contradictory.

8.44. Appropriate mathematical software produces the result $(0, 2, 4, 1)$. The rotation transforms $(0, 1, -4, 1)$ to $(0, 4, 1, 1)$, the translation transforms this to $(0, 4, 4, 1)$, and the scaling produces $(0, 2, 4, 1)$.

8.45. When T_1 or T_2 gets large, the object is magnified. However, when T_3 gets large, the object is scaled in the z direction *relative to the origin*. All the z coordinates become large, effectively moving the object away from the observer. When all three scale factors get large, the magnification in the x and y directions is canceled out by the effect of moving away in the z direction, so the object does not seem to change in size.

8.46. Equation (8.63) yields

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1/2 & 0 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix},$$

and we know that $(0, 1, -4, 1)\mathbf{T} = (0, 2, 0, 5)$. We are looking for a point $\mathbf{P} = (x, y, z)$ such that $(x, y, z, 1)\mathbf{T} = (0, 0, 0, w)$ for any $w \neq 0$. The explicit form of this set of equations is $(x, -z/2, 0, y + 4) = (0, 0, 0, w)$, and this is satisfied by all the points of the form $(0, y, 0)$, where $y \neq -4$. The interpretation of this result is simple. The rotation brings the points on the y axis to the z axis, where they are translated by three units and remain on the z axis. The scaling doesn't move these points any farther. Point $(0, -4, 0)$ is rotated to $(0, 0, -4)$ and translated to $(0, 0, -1)$, which is the viewer's position. All the points on the z axis are projected to the origin, except the viewer's location. The projection of the viewer is undefined because the case $z = -k$ results in Equation (8.59) having a zero denominator. The next example sheds more light on the perspective projection of points with negative z coordinates.

9.1. Figure Ans.30 illustrates the solution in two parts. Once seen, it looks simple!

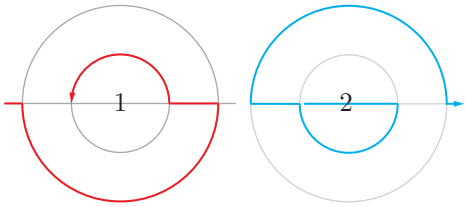
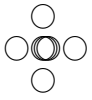


Figure Ans.30: Solution to Exercise 9.1.

9.2. Figure Ans.31 is big enough to show the empty gaps between objects. The six objects are not enough to cover the entire area and the narrow gaps add up to two squares. There is no such thing as free chocolate.

9.3. The solution, shown in the figure, requires us to stop thinking that the crossing point consists of only one coin.



9.4. Figure Ans.32 shows a spiral search pattern where in each time step we search the point with the next blue number. Such a search will find the submarine after a finite number of steps.

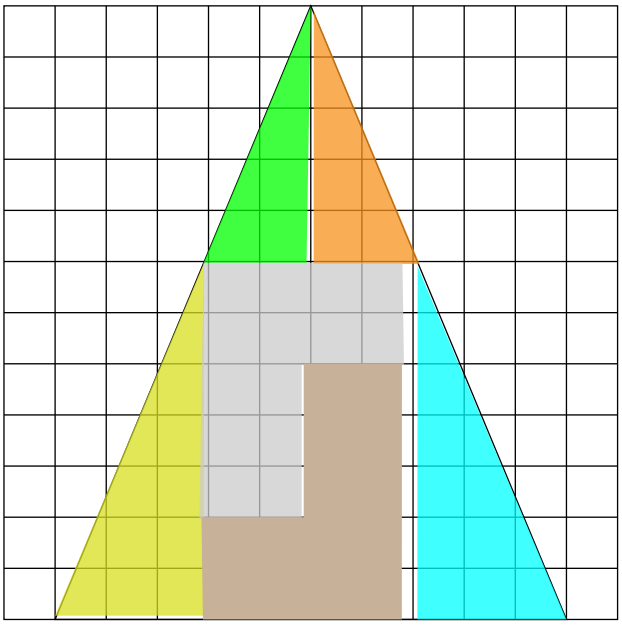


Figure Ans.31: 58 Chocolate Squares.

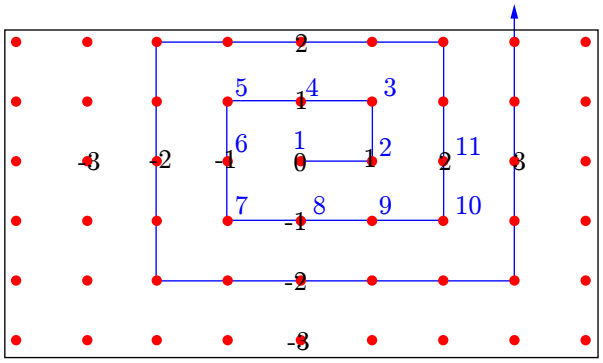


Figure Ans.32: A Spiral Search Pattern.

10.1. One example of a large factorial is 52!, the number of possible shuffles of a deck of cards. This number is so large that it vastly exceeds any estimate of the number of particles in the observable universe. In fact, it is so large that we can safely claim that your next shuffle will be unique! No one in human history has come up with this shuffle so far and no one within the life of our sun would come up with it either!

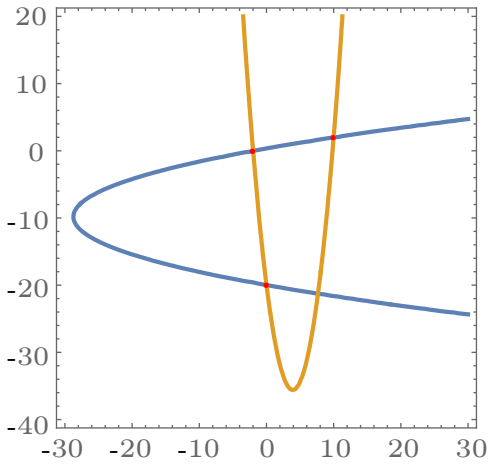
10.2. When even parity is used, the number of 1-bits in a codeword is complemented by the parity bit to an even number. If the codeword was originally all zeros, its even

parity would then be zero, causing the entire codeword to be zeros. This may not be a problem, but some software designers, especially in the past, felt that it is better to use odd parity where no codeword can be all zeros.

10.3. A general parabola can be expressed by its standard equation $c_1x^2 + c_2x + c_3y + c_4 = 0$ (for a vertical parabola) or $c_1y^2 + c_2x + c_3y + c_4 = 0$ (for a horizontal parabola). Given the three points (x_1, y_1), (x_2, y_2), and (x_3, y_3), the equation of the unique parabola passing through them is computed by solving one of the determinant equations

$$\begin{vmatrix} x^2 & x & y & 1 \\ x_1^2 & x_1 & y_1 & 1 \\ x_2^2 & x_2 & y_2 & 1 \\ x_3^2 & x_3 & y_3 & 1 \end{vmatrix} = 0, \qquad \begin{vmatrix} y^2 & x & y & 1 \\ y_1^2 & x_1 & y_1 & 1 \\ y_2^2 & x_2 & y_2 & 1 \\ y_3^2 & x_3 & y_3 & 1 \end{vmatrix} = 0.$$

Figure Ans.33 shows an example of the two orientations of parabolas. Mathematica code is included.




```
x1=-2; y1=0; x2=10; y2=2; x3=-0; y3=-20;
dv=Det[{{x^2,x,y,1},{x1^2,x1,y1,1},{x2^2,x2,y2,1},{x3^2,x3,y3,1}}]
dh=Det[{{y^2,x,y,1},{y1^2,x1,y1,1},{y2^2,x2,y2,1},{y3^2,x3,y3,1}}]
(*dv=Vertical parabola. dh=Horizontal parabola*)
gc=ContourPlot[{dh==0,dv==0},{x,-30,30},{y,-40,20}];
gp=ListPlot[{{x1,y1},{x2,y2},{x3,y3}},PlotStyle->{Red}];
Show[gc,gp]

4800 + 1912 x - 244 x^2 + 240 y
1760 + 880 x - 4792 y - 244 y^2
```

Figure Ans.33: Horizontal and Vertical Parabolas for Three Points.

10.4. We denote the ith bits of the plaintext, the keystream, and the ciphertext by d_i, k_i, and $c_i = d_i \oplus k_i$, respectively. We assume that the keystream is random, i.e.,

the probabilities $P(k_i = 0)$ and $P(k_i = 1)$ equal 0.5 each. The plaintext isn't random, so we assume that $P(d_i = 0) = p$, which implies $P(d_i = 1) = 1 - p$. Table Ans.34 summarizes the four possible cases of d_i and k_i and their probabilities. The values of c_i and their probabilities for those cases are also listed. It is easy to see from the table that the probability of c_i being 0 is $P(c_i = 0) = p/2 + (1 - p)/2 = 1/2$, and similarly $P(c_i = 1) = 1/2$. The ciphertext produced by the Vernam cipher is therefore random, which makes this simple method unbreakable.

d_i	$P(d_i)$	k_i	$P(k_i)$	c_i	$P(c_i)$
0	p	0	1/2	0	$p/2$
0	p	1	1/2	1	$p/2$
1	$1-p$	0	1/2	1	$(1-p)/2$
1	$1-p$	1	1/2	0	$(1-p)/2$

Table Ans.34: Truth Table of a Vernam Cipher.

10.5. The average word size in English is 4–5 letters. We therefore start by examining 4-letter words. There are 26 letters, so the number of combinations of four letters is $26^4 = 456{,}976$. A good English-language dictionary contains about 100,000 words. Assuming that half these words consist of four letters, the percentage of valid 4-letter words is $50000/26^4 \approx 0.11$. The percentage of five-letter words is obtained similarly as $50000/26^5 \approx 0.004$. We can therefore expect random text to contain a few short (2–4 letters) valid words, and very few 5–6-letter valid words. Longer valid words would be extremely rare.

10.6. This extra non-exercise is included because the end of the auxiliary answers file is sometimes cut off mysteriously by the typesetting software (TEXlive used in Plain TEX).

Lorem ipsum odor amet, consectetuer adipiscing elit. Leo potenti mauris imperdiet velit; interdum gravida vulputate condimentum. Mollis vestibulum senectus suscipit rutrum himenaeos at. Rhoncus fames cursus pretium nisl, pharetra ipsum morbi. Magnis vivamus nec tincidunt volutpat amet. Senectus dictum ipsum faucibus sagittis lacus cras tellus dignissim posuere. Habitasse sit scelerisque potenti enim amet. Sed rutrum ex curabitur nascetur nec egestas ultricies. Euismod ligula mus mi nisl turpis adipiscing ridiculus quam. Mauris rutrum potenti nascetur accumsan consequat duis.

Id phasellus mattis lobortis dis, ultrices maecenas. Quis class bibendum quis quis ac et nascetur. Quam urna accumsan ut senectus ultricies placerat potenti. Ipsum ligula nisl amet eu id ante. Nisl imperdiet volutpat felis fringilla proin nibh est vehicula. Dolor hendrerit nisi vestibulum senectus habitant pellentesque sem. Conubia pharetra aptent adipiscing ipsum malesuada mollis. Cras habitant torquent elit rhoncus magnis. Pretium accumsan accumsan aliquet sagittis justo porta fermentum habitasse.

Nibh in gravida turpis odio at vitae? Ex tristique vulputate at nostra curae non suspendisse ultricies. Scelerisque leo fermentum ac; ultricies diam nibh hendrerit phasellus. Velit quisque neque venenatis curabitur porta amet porta sem. Enim montes proin vestibulum maximus donec parturient natoque. Vitae eleifend molestie egestas cras dolor molestie porta. Enim ullamcorper augue pellentesque dapibus, hac diam suscipit.

Consectetur primis malesuada magna adipiscing vehicula fames dui. Potenti eu ante adipiscing proin ridiculus in ut. Consectetur nulla ullamcorper vivamus vestibulum tincidunt nascetur sollicitudin nulla. Enim maecenas eleifend pretium tortor pharetra id. Pretium fermentum consectetur phasellus sollicitudin dolor sem duis. Augue vivamus eleifend hendrerit, accumsan egestas eleifend ut habitant. Molestie magna fermentum; tellus convallis orci tortor. Etiam primis elit lobortis et condimentum sit laoreet. Massa semper nisl vehicula lacinia non mus tempus imperdiet.

Tempor ipsum nostra mollis facilisis aliquam eget eleifend dapibus nunc. Mi nunc accumsan sodales, eros nam fames accumsan. Purus taciti varius etiam luctus dapibus volutpat a habitant. Ante est tristique ligula duis vitae ligula blandit nullam. Per curae malesuada sociosqu semper accumsan vitae ante. Velit fringilla blandit curae elit dapibus? Auctor vel congue enim feugiat felis sagittis tortor porta. In bibendum accumsan eget vehicula etiam sollicitudin.

Lorem ipsum odor amet, consectetuer adipiscing elit. Leo potenti mauris imperdiet velit; interdum gravida vulputate condimentum. Mollis vestibulum senectus suscipit rutrum himenaeos at. Rhoncus fames cursus pretium nisl, pharetra ipsum morbi. Magnis vivamus nec tincidunt volutpat amet. Senectus dictum ipsum faucibus sagittis lacus cras tellus dignissim posuere. Habitasse sit scelerisque potenti enim amet. Sed rutrum ex curabitur nascetur nec egestas ultricies. Euismod ligula mus mi nisl turpis adipiscing ridiculus quam. Mauris rutrum potenti nascetur accumsan consequat duis.

Id phasellus mattis lobortis dis, ultrices maecenas. Quis class bibendum quis quis ac et nascetur. Quam urna accumsan ut senectus ultricies placerat potenti. Ipsum ligula nisl amet eu id ante. Nisl imperdiet volutpat felis fringilla proin nibh est vehicula. Dolor hendrerit nisi vestibulum senectus habitant pellentesque sem. Conubia pharetra aptent adipiscing ipsum malesuada mollis. Cras habitant torquent elit rhoncus magnis. Pretium accumsan accumsan aliquet sagittis justo porta fermentum habitasse.

Nibh in gravida turpis o

I don't have any solution, but I certainly admire the problem.

—Ashleigh Brilliant.



Index

The index is detailed and constitutes about 4% of the book. This was done even though the book was published in PDF, which allows for quick online search, because of the following:
1. Some items may not be easy to locate in an online search. A two-word item such as `bilinear interpolation`, may not appear in a PDF search if it happens to be broken across two lines.
2. Certain readers may decide to print the book, in spite of its large size.

I have also attempted to make the index items as complete as possible, including middle names and dates. Any errors and omissions brought to my attention are welcome. They will be added to the errata list and will be included in any future editions.

(•) point (in orbifold notation), 190
0^0 (valid?), 267–269, 272
$\sqrt{2}$
  continued fraction expansion of, 86, 90, 94, 639
  is proved irrational, 4–6, 82
  very irrational, 90
× miracle (in orbifold notation), 191
(1,2)-Fibonacci sequence, 287
* (orbifold reflection indicator), 190
$0.9\overline{9} = 1$ proof, 6–7, 100, 271, 272
14-15 puzzle, 553
15 puzzle, 576
16-room puzzle, 553
17 wallpaper symmetry types, 196–197
196 (boring?), 68
196 Palindrome Quest, 234
2, 3, 4, 82,000 Sequence, 284
264 magic square, 583
$\sqrt[3]{2}$ is irrational, xxii
3-Fibonacci sequence, 287
37 (not random?), 63, 356
4 (as an attractor), 55
4-digit pins (choosing of), 383–385
7 (as a base of a numbering system), 663
9 (as an attractor), 60

A

Abbott, William Alexander "Bud" (1897–1974), 380
Abel, Niels Henrik (1802–1829), 277
Abelian group, 172
absolute convergence (of series), 304, 306
abundant number, 71, 75, 283
Achilles and the fly paradox, 249, 325
achirality (opposite of chirality), 169
Ackermann, Wilhelm Friedrich (1896–1962), 48
Adler, Alfred W. (1930–), 29
Aerial (song), 141

affine transformations, 425–426
AGM, arithmetic-geometric mean function, 137
Agnesi, Maria Gaetana (1718–1799), 35
Ahmes (ancient scribe), 645
al-Khwârizmî, abu Ja'far Muḥammad ibn Mûsâ (c. 780–850), 80
Albert of Saxony (1316–1390), 243
Alberti, Leon Battista (1404–1472), 520–527, 684
$\aleph_0$ (pronunciation of), 253
algebra (fundamental theorem of), 103
algebraic numbers, 71, 256
algorithmic information theory, 100, 359
aliquot sequence, 64, 283–284
aliquot sum, 75, 76, 283
Allen, Tim (1953–), 46
almost integer numbers, 71
alternating harmonic series, 304, 306
Altmann, Simon L., 124
ambigram (definition of), 233
amicable numbers, 71, 283
Amis, Kingsley William (1922–1995), 322
analytic geometry, 4, 20
Anaxagoras of Clazomenae (c.500–c.428 B.C.), 8
Anaximander, (c.610–c.546 B.C.), 240
Anderson, Chris (1957–), 416
anti-Fibonacci numbers, 286
anti-magic squares, 588
antikythera (ancient orrery), 95
antinomy
  definition of, 318
  example of, 330
antipalindromic number, 234
antipodes (definition of), 612
antiprime (highly composite) numbers, 73
aperiodic tiling, 222–228, *see also* nonperiodic tilings
apocalypse numbers, 76
apocalyptic power, 76
apothem (of a regular polygon), 548
Appel, Kenneth Ira (1932–2013), xxi
approximating curve, 503
arbelos (geometric configuration), 232
Archimedean solids, 377, 378
Archimedean spiral, 28
Archimedean tiling, *see* semiregular tilings
Archimedes of Syracuse (287–212 B.C.), 135, 140, 232, 377
area of a regular polygon, 548
Argand diagram, 105, 144, 641
  and conjugate, 105
  and roots of unity, 615
  and rotations, 121
  and trinions, 112, 113
  introduction of, 106
Argand, Jean-Robert (1768–1822), 105, 106
Aristotle (384–322 B.C.), 261, 320, 331, 617
Aristotle's wheel paradox, 330
arithmetic average, 137, 607
  and expectation, 401
  weighted, 349
arithmetic operations, 47–49
arithmetic, fundamental theorem of, 73
aspiring number, 283
asymmetric object (nonesuch?), 177
asymmetry (definition of), 168
Atanasoff, John Vincent (1903–1995), 635
Atiyah, Michael Francis (1929–2019), xviii
atmospheric perspective, 520
attractor
  the digit **4**, 55
  the digit **9**, 60
attractor (definition of), 416
Austen, Jane (1775–1817), 633
Austin, A. K., 325
Austin, Tavon, (1990–), 453
autological word (definition of), 343
average
  geometric, 407
  harmonic, 406
  types of, 607–608
  weighted, 607
average die, 652
Avogadro's number, 346
axiom of choice, 9–12
axis (in geometry), 457
  and vectors, 457

B

Baez, John Carlos (1961–), 59
Bagehot, Walter (1826–1877), 353
Balzac, Honoré de (1799–1850), 477
Banach-Tarski paradox, 11
Bandura, Albert (1925–2021), xxxiv
barber paradox, 321

Barnsley, Michael Fielding (1946–), 415
Barrow, Isaac (1630–77), 263
barycentric (meaning of), 471
barycentric sum, 401, 470
barycentric weights, 486
base of a numbering system, 288
  any?, 284, 288, 663
base-φ numbers, 288
base-10 numbers, 49, 52, 60, 288
base-13 (Conway function), 103, 433
base-16 numbers, 63
base-2 numbers, 49, 52, 608
  length of, 53
base-8 numbers, 58
Basel problem, 314
Bayes' rule, 339, 340, 666
Bayes, Thomas (1701–1761), 339
beastly primes, 65
beauty in mathematics, xvi, 2, 4, 79, 80, 139
Becquerel, Antoine Henri (1852–1908), xx
Bell, Eric Temple (1883–1960), 256, 342
Belphegor number, 65
Benford's Law, 361–376
  and Excel, 374
Benford's law, 376
Benford, Frank Albert (1883–1948), 361, 362
Berger, Robert (1938–), 227
Bergfeldt function, 432
Bergfeldt, Axel, 432
Berlekamp, Elwyn (1940–2019), 603
Bernoulli trial, 349
Bernoulli's spiral, *see* equiangular spiral
Bernoulli, Daniel (1700–1782), 323, 582
Bernoulli, Jacob (1655–1705), 34, 131, 350
Bernoulli, Johann (1667–1748), 474
Bernshteĭn, Sergeĭ Natanovich (1880–1968), 506
Bernstein polynomials, 507
  in *Mathematica*, 507
Bernstein, Peter Lewyn (1919–2009), xxii
Berra, Lawrence Peter (1925–2015), 314
Berry's Paradox, 342, 636
Berry, Clifford Edward (1918–1963), 635
Bertrand box paradox, 339–340, 666
Bertrand paradox, 337–339
Bertrand, Louis François (1822–1900), 337, 339
Bézier curve, 503–510, 543
  Bernstein form of, 506–509
  collinear points, 510, 682, 684
  control polygon, 504
  cubic, 509
  degree 4, 510
  heart shaped, 508, 680
  in *Mathematica*, 507
  parameter substitution, 510, 681
  quadratic, 509
Bézier methods, 503–510
Bézier, Pierre Etienne (1910–1999), 503, 506
Bhaskara-Brouncker algorithm, 663
bigram, *see* digram
bijection, *see* bijective function
bijective function (definition of), 173
bilinear interpolation, 435, 445–448
  smoothed, 447–448
bimagic squares, 585
binary numbers, *see* base-2
Binet formula, 156, 285, 286
Binet, Jacques Philippe Marie (1786–1856), 156, 286
binomial coefficient, 505, 507
binomial expansion theorem, 308
binomial theorem, 504–506
birthday problem, 16–18, 385–386
black swan theory, 351
Blake, William (1757–1827), 274
Bleckley, Logan E. (1827–1907), 349
blending
  Bernstein polynomials, 507
  Hermite, 483–487
  Hermite derivatives, 488–489
  Hermite functions, 486–487
  parabolic, 481, 499–502
  in a PC, 476
block-stacking problem, 312
Bloom, Molly (fictional character), 611
Bogomolny, Alexander (1948–2018), 8
Bohr, Niels David (1885–1962), 344
Bolzano, Bernardus Placidus Johann Nepomuk (1781–1848), 266, 331, 430
Bombelli, Rafael (1526?–1572), 104
book-stacking problem, 312
boring numbers, 2–3, 68, 71, *see also* rarely used numbers
Borsuk-Ulam theorem, 612
Bortkiewicz, Ladislaus Josephovich (1868–1931), 667

bounding box (square?), 613
Boyle, Charles (1674–1731), 94
braccio (unit of length used in the Renaissance, plural braccia), 522
brachistochrone, 24, *see also* curve of shortest time
Brahmagupta (c. 598–c. 668), 520
Brewster, David (1781–1868), 188
brick wall (symmetries of), 653
Brilliant, Ashleigh Ellwood (1933–), 690
Brouwer fixed-point theorem, 41, 611–612
Brouwer, Luitzen Egbertus Jan (1881–1966), 611
brown numbers, 71
Brownian motion, 355, 358
  as a random process, 357
  not true random, 358
Brunelleschi's peepshow experiment, 513–517
Brunelleschi, Filippo di ser Brunellesco di Lippo Lapi (1377–1446), 512–517, 520
Buffon needle (for π), 136
Bunyan, John (1628–1688), (Colophon)
Buonarroti, Michelangelo di Lodovico Simoni (1475–1564), 513
Burnett, Carol (1933–), 579
burning ship fractal, 420
Bush, Catherine (1958–), 141
byte (definition of), 58

C

cable guy paradox, 321
Caesar, Gaius Julius (100–44 B.C.), 324
CAGD (computer aided geometric design), 503
Caine, Michael (1933–), 646
camera obscura (and left-handedness), 669
Campbell, Joseph John (1904–1987), 551
Camus, Albert (1913–1960), 360
candidate (choosing), 387–389
Cantor (diagonal argument), 255, 322
Cantor paradox, 258
Cantor sets, 414, 415
Cantor, Georg Ferdinand Ludwig Philipp (1845–1918), xiii, 10, 11, 98, 99, 129, 240, 241, 251, 252, 414
card draw, 333
Cardano, Gerolamo (1501–1576), 104, 331
cardinal splines, 481, 493, 495–498
cardinality
  definition of, 255, 403
  of power sets, 257–258
  of the algebraics, 256
  of the irrationals, 255, 256
  of the reals, 256
  of the transcendentals, 256
Carpenter, Loren C. (1947–), 415
Carroll, Lewis (1832–1898), 47, 55
Carson, John William (1925–2005), xxix
Cassini oval, 33–34, *see also* ellipse and superellipse
Cassini, Giovanni Domenico (1625–1712), 33
Catalan-Dixon conjecture, 283
Cataldi, Pietro Antonio (1548–1626), 88
Catmull-Rom curves, 497–502
Cauchy, Augustin-Louis (1789–1857), 105, 266, 430
Cauchy–Schwarz inequality, 391
causal determinism
  and Laplace, 352
  definition of, 352
Cavalieri, Bonaventura Francesco (1598–1647), 261
Cayley table, 171, 178, 652, 653
Cayley's theorem, 180
CD (compact disc), 596–598
  error correction in, 598–604
Cesàro, Ernesto (1859–1906), 415
Chaitin's constant, 71, 100
Chaitin, Gregory John (1947–), 71, 100, 359
Champernowne constant, 71, 80, 100, *see also* lexicon
Champernowne, David Gawen (1912–2000), 71, 100
chance (what is it?), 357
chaos game (fractals), 420
chaos theory, 357, 417
ChatGPT in this book, vii, xi, xii, xxiii, xxix, 16, 76, 129, 130, 162, 166, 170, 237, 274, 352, 408, 411, 422, 578, 590, 652, 665, 671
Child, Lee (James Dover Grant, 1954–), xxi, 6
chirality
  definition of, 168
  in a die, 181

in symmetry, 168
choosing a candidate, 387–389
chords (products of), 615–617
Christie, Agatha Mary Clarissa (1890–1976), 8, 188, 265, 667
Chrystal, George (1851–1911), 149
circle, 216
bounding box of, 614
decomposing, 9
dissecting, 9
scaled to ellipse, 216, 656
squaring, 8–9, 11, 135
circle inversions, 33, 230–232
circle squarer (derogatory term), 135
Clancy, Thomas Leo Jr. (1947–2013), 557
Clarke, Arthur Charles (1917–2008), 55, 341, 652
Clawson, Calvin Clarence, xxiii
clock hands overlap (puzzle), 562
codes
(n, k), 594
error-control, 591–604
Hamming distance, 594
parity, 593
Reed–Solomon, 598–604
codeword (definition of), 593
Cohen, Paul Joseph (1934–2007), 11
coin landing on its edge, 353, 354
coin puzzle, 555
coin rotation paradox, 324
coin toss, 333, 353, 355, 406, 669
coins shapes, 614
collinear points
and Bézier curve, 510, 682, 684
not independent, 470
uniformly spaced, 683
colossally abundant numbers, 75
comma sequence, 301–302
complete randomness does not exist, 15
complex golden ratio, 156–157
complex numbers, 103–114, 142–146, *see also* hypercomplex numbers
and points, 113–114
conjugate, 105
exponentiation rule, 270
fractional powers of, 109, 641
fun with, 114
history of, 104–105
rotation of, 250
composite number, 75, *see also* antiprime
computable numbers, 71, *see also* non-computable numbers
conditional convergence (of series), 305
conditional probability, 339–341, 666
conical helix, 29
Conrad, Keith (1970–), 172
constants, *see* mathematical constants
constructible numbers, 71
continued fractions, 85–97, 241, 301
and irrational numbers, 89
and Khinchin's constant, 62, 96, 638
and palindromes, 235
and quadratic equations, 86, 87
canonical form, 85
confusing, 90
definition of, 85
Euler's formula, 92–93
introduced by John Wallis, 265
of π, 87, 88
of φ, 86, 88, 90
of e, 640
of an integer, 86
of irrationals, 85
of rationals, 85
recurrence formulas for convergents, 88
continuity (of functions), 429–434
continuous function (definition of), 427
continuum
baffling, 99
no individual steps, 333
no successor, 99, 333
probabilities in, 332–335
real numbers, 4
continuum hypothesis, 256
control points, 503, 508
and convex hull, 508
reversing, 508
control points (for curve design), 476
control polygon of the Bézier curve, 504
Conway, John Horton (1937–2020), 4, 69, 186, 187, 189, 303
Copeland-Erdős constant, 637
CORDIC rotations, 464–469
cosh ("kosh," definition of), 144
countable
algebraic, 98
definition of, 79, 244

rationals, 79, 98, 99
Coward, Noel Peirce (1899–1973), xi
Coxeter, Harold Scott MacDonald (1907–2003), 150
Cristillo, Louis Francis (1906–1959), 380
cross-product, 451–453, 472, 673
crunode (definition of), 34
Crystal, William Edward (1948–), 646
cubic Bézier curve, 509
cubic polynomials, 474–477
cubic splines, 481, 490–494
  clamped, 493
  indefinite direction, 494
cuboctahedron, 378
Curd, John, 63
curve of shortest time (cycloid), 23
curves
  Bézier, 503–510, 543
    cubic, 509
    degree 4, 510
    parameter substitution, 510, 681
    quadratic, 509
  Cassini oval, 33–34
  Catmull-Rom, 497–502
  cycloid, 23
  direction vector, 483
  egg shape, 32–33
  explicit representation of, 472
  heart, 37–38
  helices, 28–29
  Hermite, 483–489
    indefinite tangent vectors, 679
  hypotrochoids, 22–23
  implicit representation of, 472
  inflection points, 475
  parabola, 26–27
  parametric representation of, 473
  representations of, 472–473
  slope of, 474
  spirals, 28–29
  superellipse, 35–37
  tangent vector, 476, 477, 481, 488
    Bézier, 508
    definition of, 473, 474
    direction of, 676
    extreme, 491, 493
    indeterminate, 489, 683, 684
  tension in, 493, 495–498
  velocity of, 477
  witch of Agnesi, 35
curves and surfaces, 20–38
curves of constant width, 614–615
cusp, 503, 683, 684
cycle notation of permutations, 180
cyclic group
  definition of, 177
  isomorphic, 177
cyclic numbers, 66, 71, 164–165
cycloid, *see also* brachistochrone
  curve of shortest time, 23
  definition of, 23

D

D_{10} symmetry group, 178
D_4 dihedral group, 176
D_8 symmetry group, 178
Dali, Salvador (1904–1989), 596
dartboard paradox, 332–335
Darwin, Charles Robert (1809–1882), xx, 359
data
  qualitative, 363
  quantitative, 363
data points (for curve design), 475
da Vinci, Leonardo (1452–1519), 513, 520
de Moivre formula, 105, 109
de Moivre, Abraham (1667–1754), 105, 109
De Morgan, Augustus (1806–1871), 577
decimal numbers, *see* base-10
decompositions of the sphere, 11
defective number, *see* deficient number
deficient numbers, 71, 75, 76, 283
definable numbers, 71
deformations
  successive rotations, 463
del Ferro, Scipione (1465–1526), 104
delayed palindrome, 234
Democritus, (c. 460–c. 370 B.C.), 261
denumerable, *see* countable
Descartes, René (1596–1650), xii, xxii, 4, 103, 643
determinant
  and cross product, 452
  pure reflection, 218
  pure rotation, 218
  use in scaling, 216
determinism (definition of), 352

Diaconis, Persi (1945–), 383
diagonal argument (Cantor), 255, 322
diagonally-dominant matrix, 493
dialetheia (definition of), 318
dice
  go first, 380–382
  randomness in, 376–380
  symmetries of, 180
differentiable function (definition of), 427
digamma function $\psi(x)$, 312
digram (definition of), 359
dihedral group D_4, 176, 180
dihedral group $D_{2\cdot 3}$, 179
dihedral group $D_{2\cdot 4}$, 178, 180
dihedral group $D_{2\cdot 5}$, 178
dihedral group (definition of), 176
dimension (meaning of), 328
dimensions, non-commensurate, 665
Diophantus of Alexandria, xxii
Dirac, Paul Adrien Maurice (1902–1984), 279
direction cosines, 459
direction vector of Hermite curve, 483
Dirichlet function, 431, 432
  everywhere discontinuous, 431
Dirichlet, Johann Peter Gustav Lejeune (1805–1859), 431
disagreements in mathematics, xix, 279, 382
discontinuous function (definition of), 427
discontinuous functions (everywhere), 430–434
dissecting a square, 552
divergent series, 54
divided differences, 481
divine proportion, *see* golden ratio
divine section, *see* golden ratio
divisibility tests, 54, 55, 57, 59, 60, 635, 636
dodecahedron (as a kaleidoscope), 188
Dostoevsky, Fyodor Mikhailovich (1821–1881), xxiii
dot product, 221, 472
  and Perlin noise, 444
double factorial, 279
double magic square, 585
double miracle symmetry, 195
dungeon numbers, 71
Dunham, William Wade (1947–), 252
duodecimal number system, 63
duonions (complex numbers)?, 114
Dürer, Albrecht (1471–1528), 527, 583

E

e, Euler constant, 62, 81, 270
  as natural base, 132
  continued fraction of, 640
  transcendental, 248
Eddington, Arthur Stanley (1882–1944), 277
egg shape, 32–33, 634
Einstein tile, 226–227
Einstein, Albert (1879–1955), xiv, xix, 2, 227, 240, 410, 556, 557
  birthday, 136
  play dice?, 355
Eliot, George (1819–1880), 389, 551
ellipse, 216
  and billiard, 32, 634
  circle inversion of, 33
Ellsberg paradox, 383
Ellsberg, Daniel (1931–2023), 383
enneanacci numbers, 287
equations, 275–280
equiangular spiral, 28, 160
equilateral triangle (symmetries of), 178, 180
Erdős, Paul (1913–1996), 24, 162
error-control codes, 591–604
error-correcting codes in a CD, 598–604
escape-time (fractals), 419–420
Escher's horse, 658, 659
Escher's lizard, 230
Escher, Maurits Cornelis (1898–1972), 170, 415, 535, 537
Estermann, Theodor (1902–1991), 83
ETAOIN SHRDLU (letter distribution), 358
eternal fame (in mathematics), 2
Euclid, (ca. 325–270 B.C.), 149, 151, 335, 474
  and ϕ, 157
  axioms, 11
  GCD method, 79, 88
Euler angles, 455
Euler equation, xvi, 105, 106, 109, 136, 250, 270, 648, 661
  and quaternions, 123–127
  contains π, 135
Euler prime generator function, xx

Euler series, 314, 638
  proof of, 314
Euler Trail puzzle, 555
Euler's constant γ, 130
Euler's continued fraction formula, 92–93
Euler's number e, 130–134, 163, 293, 308, 638
  as natural base, 132
Euler, Leonhard (1707–1783), 62, 88, 92, 105, 106, 109, 123, 130, 250, 314, 455, 555, 556, 582, 636, 638
everywhere discontinuous functions, 430–434, *see also* nowhere differentiable function
everywhere nondifferentiable function, 427
Excel spreadsheet, 374
exclusive-OR (XOR), 49, 608–611
  in Vernam cipher, 611, 689
existence (in mathematics), xiii
expected value
  of a dice roll, 349
  of a random experiment, 349
explicit representation of curves, 472
extended Fibonacci sequence, 161, 296
extreme and mean ratio, *see* golden ratio
extrinsic rotations, 455

F

Faber, Georg (1877–1966), 415
factorial (double), 279
factorion, 71
  definition of, 71
fair by continuity, 379
  polyhedra, 379
fair by symmetry (polyhedra), 378
fair game (definition of), 323, 324
Falbo, Clement E. (1935–), 150
falsidical paradox, 317
Farber, David J. (1934–), 599
Fatou, Pierre Joseph Louis (1878–1929), 414
Faucher, Kane X., 273
favorite numbers, xxviii–xxix
Feigenbaum constant, 81
Ferguson, James C., 490
Fermat's Last Theorem, xxi, xxii, 633
Fermat, Pierre de (1607–1665), xxii, 4, 263, 265
Fewster, Rachel Mary (1974–), 374
Feynman, Richard Phillips (1918–1988), 391, 392
fiasco (definition of), 405
Fibonacci integer sequences, 286
Fibonacci numbers, 52, 71, *see also* anti-Fibonacci numbers
  and ϕ, 153
  and 91, xxviii
  and L-systems, 418
  and powers of ϕ, 152, 296
Fibonacci primes, 285
Fibonacci sequence, 161, 285–296, 300, *see also* extended Fibonacci sequence
  a base of a numbering system, 288, 289
  and Pisano period, 287
  as truncated continued fraction, 301
  converting between miles and kilometers, 288
  increasing monotonic, 75
  odd and even numbers, 366
Fibonacci variable-length code, 288–291
Fibonacci word, 287
Fibonacci, Leonardo Pisano (1170–1250), 52, *see also* Leonardo of Pisa
Fibonacci-esque numbers, 301
fifth roots of unity, 109–110
figurate numbers, 53
finding a birthmate, 386
fine-structure constant, xxviii
finite field, 602
five-body system, 337
fixed-point theorem (Brouwer), 41, 611–612
flag of Nepal, 618
foci (pronunciation of), 32
Foias constant, 81
Foucher, Simon (1644–1696), 246
four color map theorem, xxi, 55
four-square theorem, 55, 57
fractal dimension, 411, 412
fractals, 410–420
  binary tree, 419
  burning ship, 420
  Cantor sets, 414, 415
  carpet, 240, 414, 422
  chaos game, 420
  definition of, 410, 411
  dragon curve, 240, 415
  escape-time, 419–420
  fragmented circle, 659

Haferman carpet, 415
history of, 414–415
iterated function systems, 415–416, 421–426
Julia, 240, 414, 420
Koch, 328, 412, 414, 415, 425, 665
L-systems, 418–419, 454
Levy C curve, 414
Levy curve, 424
logistic map, 417
Lyapunov, 420
Mandelbrot set, 420
Menger sponge, 240, 415
nova fractal, 420
Peano curve, 415
self-inverse, 414
Sierpinski, 412
Sierpinski hexagon, 424
Sierpinski square, 424
Sierpinski triangle, 414, 415, 421, 424
Sierpinski's secret (poem), 422
strange attractors, 416–418
T-square, 415
Weierstrass function, 414
fractional powers complex numbers, 109, 641
Fraenkel, Abraham (1891–1965), 10
Francesca, Piero Della (1415–1492), 32, 34, 521
Fransén–Robinson constant, 81
free chocolate (no such thing), 686
frieze patterns, 190
fun with complex numbers, 114
functions
Bergfeldt, 432
bijective, 173
continuity of, 429–434
continuous, 427
Conway base 13, 433
definition of, 427
differentiable, 427
digamma, 312
Dirichlet, 431, 432
discontinuous, 427
everywhere discontinuous, 430–434
everywhere nondifferentiable, 427
Gamma, 312, 581–582, 647
general, 173
Heaviside, 427
injective, 173
main types of, 173
nondifferentiable, 427
nowhere differentiable, 427
orthogonal, 221
Signum, 427
surjective, 173, 433
Thomae, 432
Weierstrass, 414, 427–429
fundamental region of a pattern, 193
fundamental theorem of algebra, 103

G

Gabriel's horn, 327
Galileo di Vincenzo Bonaiuti de' Galilei (1564–1642), xiii, xvi, 246, 261, 262, 320, 331, 347
Galileo's paradox, 320
Galois field, 602
Galois, Évariste (1811–1832), xiii, 170
gambler's fallacy, 350
game (fair?), 323, 324
Gamma function, 312, 581–582
of i, 647
Gardner, Martin (1914–2010), 325, 658
Gaudi, Antoni i Cornet (1852–1926), 583
Gauss, Johann Carl Friedrich (1777–1855), xxviii, 103, 137, 257
and i, 142
and imaginary quantities, 103, 644
Gauss-Legendre formula for π, 137–138
Gelfond–Schneider constant, 81
general function (definition of), 173
generalized golden spiral, 160
generator of a cyclic group, 177
geometric average, 407
geometric mean, 137, 607
Givens rotations, 461–462
Givens, James Wallace Jr. (1910–1993), 463
glide reflection, 186
miracle, 195
go-first dice, 380–382
God (love of), 637
Godard, Jean Luc (1930–2022), 390
Gödel, Kurt Friedrich (1906–1978), 11
Goethe, Johann Wolfgang von (1749–1832), 130, 169
Gogh, Vincent Van (1853–1890), 527
golden angle, 64, 81, 147, 291

Golden Ratio, 468
golden ratio, 61, 130, 147–152, 285, 300, 637, 638, *see also* Fibonacci numbers, Lucas sequence
  a contrary viewpoint, 150
  and golden angle, 64, 291
  and Penrose tiles, 226
  and sunflower patterns, 295
  complex, 156–157
  continued fraction expansion of, 86, 88, 90
  generalization of, 150, 157–161
  history of, 149–151
  most irrational number, 90, 150, 293, 295
  rational approximation of, 94
  unusual appearances of, 153–154, 418
golden rectangle, 152, 299, 649
  self similar, 160
golden spiral, 152, 299
  generalized, 160
  self similar, 160
Goliath number, 76
Golomb sequence, 283
Goodwin, Edward J. (c.1825–1902), 645
Grabowski, Łukasz, 9
gradient (meaning of), 443
Graham's number, xxix
Graham, Ronald Lewis (1935–2020), xxix
Grandi's series, 246, 249
Granville numbers, 71
graphics (visible math objects), 19–41
Graves, John Thomas (1806–1870), 59, 117, 644
Grelling, Kurt (1886–1942), 343
Grey, Thomas (1455–1501), 235
Griffin, Grant R., 469
group
  and symmetry, 170
  definition of, 172–173
groups
  Abelian, 172
  cyclic, 177
  dihedral, 176
  monster, xxviii, 171
  order of an element, 172
  rotation, 450
  symmetry, 172–186
  trivial, 176
Guy, Richard Kenneth (1916–2020), 282, 667
gyration
  definition of, 187
  example, 187, 192
  in orbifold signatures, 192
  order of, 187

H

Hadamard, Jacques Salomon (1865–1963), 483
Haferman carpet, 415
Haig, Matt (1975–), xix
Haken, Wolfgang (1928–2022), xxi
Hales–Jewett theorem, 14, 15
Hall (Halparin), Monte (1921–2017), 341
Hals, Frans (1580–1666), 511
Hamilton, William Rowan (1805–1865), 59, 113, 115, 116, 118, 126, 128, 644
Hamming distance, 594
Hamming, Richard Wesley (1915–1998), xviii, 594
Hardy, Darel, 506
Hardy, Godfrey Harold (1877–1947), xiii, 65, 430
Hardy-Ramanujan number, 65
harmonic average, 406
harmonic mean, 607
  in HLL, 405
harmonic numbers, 309
harmonic series, 307, 309–314, 432
  diverges, 51
  real, 309–312
Harry Potter (novels), 58
Harshbarger, Eric C., 380
Harter-Heighway dragon curve, 240, 415
hashing and hash tables, 404–405
hat, *see* Einstein tile
Hawking, Stephen William (1948–2018), xiv, 510, 646
headword (definition of), 592
heart shape, 37–38, 42, 379
  Bézier curve, 508, 680
Heaviside function, 427
Heaviside, Oliver (1850–1925), 119, 121
Hein, Piet (1905–1996), 37, 658
Heisenberg's Uncertainty Principle, 141
helices, 28–29
helix (definition of), 29
heptagonal numbers, 58

heptanacci numbers, 287
Heraclitus, (5th–6th century B.C.), 240
Hermite blending, 483–487
Hermite blending functions, 447, 486–487
Hermite interpolation, 483–489
  derivatives, 488–489, 677
  indefinite tangent vectors, 679
  midpoint, 486, 677
Hermite, Charles (1822–1901), 130, 248, 429, 482, 483
Hernandez, Laurie (2000–), 549
Hertz, Heinrich Rudolf (1857–1894), xvii
heterological word (definition of), 343
hexadecimal numbers, *see* base-16
hexagonal numbers, 57
hexanacci numbers, 287
highly abundant number, 75
highly composite number, 75
  (definition of), 73
Hilbert curve, 235, 236, 412, *see also* space filling curves
Hilbert, David (1862–1943), 244, 251
history of fractals, 414–415
history of perspective, 511–527
HLL algorithm, 403–407
Hobbes, Thomas (1588–1679), 265
Hockney, David (1937–), 510, 511, 527
Hoffman, Paul (1953–), 5
Hofstadter, Douglas Richard (1945–), 12
homogeneous coordinates, 545
homological word, *see* autological word
Hooke, Robert (1635–1703), 520
horizon line
  high, 532
  low, 535
  placement of, 521, 532–537
horror paradox, 324
horse (Escher), 658, 659
Householder, Alston Scott (1904–1993), 463
Hubble, Edwin Powell (1889–1953), xvii
Hume, David (1711–1776), 393
humor in this book, 2, *see also* jokes in this book
Humphrey, Hubert Horatio (1911–1978), 316
husbands (jealous), 574
Huxley, Aldous Leonard (1894–1963), 243
hyper-4, *see* tetration
hyperbolic functions, 144
hypercentric camera, 527
hypercomplex numbers, 71, 112–115
hyperoperation sequence, 47–49
hyperperfect numbers, 71
hyperreal numbers, 71
hypotrochoid (curve), 22–23

I

i, ($\sqrt{-1}$), 142–146
  (origin of), 103, 142
  why this notation, 142
I hate mathematics, x
icosidodecahedron, 378
iff (meaning of), 173
IFS, iterated function systems, 415–416, 421–426
Ignatius of Loyola (1491–1556), 264
i^i is real, 144
illegal numbers, 71
image by ChatGPT, xi, 16
image by MS Copilot, 3, 8, 51, 60, 65, 95, 126, 252, 313, 318, 326, 329, 342, 350, 354, 388, 416, 559, 565, 567, 571, 576, 662, 664, 669, 672, 692
image compression (SVD), 604–606
imaginary numbers, 103–104
  origin of, 643
immediate successor (none), 4, 79, 99, 427
implicit representation of curves, 472
improper rotation, 187
infinite
  mysterious nature of, 51, 130, 240–242
  paradoxical nature of, 244, 317, 320, 328, 333, 334
infinite gift (and Gabriel's horn), 665
infinite grand hotel, 244–246
infinitesimals (history of), 261–266
infinity, 239–274
  and human stupidity, 2
  as a limit, 52
  in math, xiii, 6–7
  not a number, 52, 267
  unexpected results of, 52, 247–251, 659
infinity in finite time (paradox), 335–337
inflection point
  definition of, 154
  of a quartic polynomial, 154, 155

inflection points, 475
  in a PC, 475
injective function (definition of), 173
integers, 50–73
  continued fractions of, 86
  infinitely many, 3
  square roots are irrational, 81
  types of, 75
integral (definite) must have a constant, 271
interpolation
  Hermite, 483–489
  Lagrange, 478–480
  Newton, 480–481
  spline, 481–503
interpolation methods, 434–510
  bilinear, 435, 446–447
  linear, 477
  points and vectors, 469–472
intrinsic rotations, 455
irrational numbers, 78–97, 639
  and broken periodicity, 38
  and randomness, 355
  definition of, 78
  lots are known, 81, 82
irrationality, 82, 83, 93
  measuring of, 89
  of the transcendence of a number, 135
isocephaly (meaning of), 521
isometric (objects with the same symmetries), 177
isometry, 171
  definition of, 186
  improper rotation, 187
  rigid motion, 186
isomorphic (meaning of), 177
isomorphism
  and Cayley's theorem, 174
  definition of, 174
  of $D_{2\cdot 4}$, 178
isosceles triangle (symmetries of), 179
Izmailovo hotel, 244

J

Jacobsthal sequence, *see* (1,2)-Fibonacci sequence
Jeandel, Emmanuel, 227
Jeans, James Hopwood (1877–1946), 137
Jenny's number, xxviii
jokes in this book, xiv, xxiv, 69, 102, 233, 296, 304, 328, 351, 376, 392, 395, 418, 633, 667
Jones, William, (1675–1749), 135
Joyce, James Augustine Aloysius (1882–1941), x, 233, 611
juggler sequence, 281
Julia set, 240, 414, 420
Julia, Gaston Maurice (1893–1978), 240, 414

K

Kac, Mark (1914–1984), 357
kaleidoscope, 187
kaleidoscopic point, 187, 188, 191
Kamermans, Mike, 503
Kaprekar process, xxix, 64
Kaprekar, Dattatreya Ramachandra (1905–1986), xxix, 64
Kauffman, Stuart Alan (1939–), 292
Keats, John (1795–1821), vii, 2
Keith number, 287
Keith, Michael (1955–), 287
Kepler, Johannes (1571–1630), 149, 151
  and ϕ, 286
Khinchin constant, 7, 62, 81, 96, 607, 638
Khinchin, Aleksandr Yakovlevich (1894–1959), 62
Kimball, Spencer Woolley (1895–1985), xvi
kissing numbers, 71
Klein, Felix Christian (1849 –1925), 414
knight's tour, 566, 576
  chess problem, 586
knots (in curve design), 478
knowing (or not), xv
Knuth, Donald Ervin (1938–), xxix, 48, 66, 140, 362, 379, (Colophon)
Koch curve, 425
Koch snowflake, 328, 412, 414, 415, 425, 665
König Gyula (Julius, 1849–1913), 322
König's paradox, 322
Königsberg bridge problem, 555
Kovalevskaya, Sofia (1850–1891), 280
Kronecker, Leopold (1823–1891), 71, 251, 257

L

L-systems (fractals), 418–419, 454
Lévy 2 constant, 81

Lévy, Paul Pierre (1886–1971), 414
Laan, Hans van der (1904–1991), 299
Laczkovich, Miklós (1948–), 9
Lagrange interpolation, 478–480
Lagrange polynomial, 478–480, 581, 675
  quadratic, 479
Lagrange's theorem, 182
Lagrange, Joseph Louis (1736–1813), 57, 173
Lagrangian points, 56
Lamé, Gabriel (1795–1870), 35
Lambert Johann Heinrich (1728–1777), 135
Lanier, Susie, 66
Laplace, Pierre-Simon de (1749–1827), 352
large numbers (notation for), xxix, 48–49
law of cosines, 278
law of large numbers, 350, 667
  and die tosses, 402
law of small numbers, 351
law of small numbers (book), 667
law of truly large numbers, 350–351
leap years, 573
least squares interpolation, 413, 435–439, 477
left handed (chance of being), 392, 669
left-hand rule (in rotation), 453
left-truncatable primes, 74, *see also* right-truncatable primes
Lehmer, Derrick Henry (1905–1991), 284
Lehrer, Thomas Andrew (1928–), 257
Leibniz, Gottfried Wilhelm (1646–1716), xii, 52, 246, 263, 265, 414
Lemniscate constant, 81
lemniscate of Bernoulli, 34
length of an integer, 406
Leonardo of Pisa (Fibonacci, c. 1170–c. 1250), 151, 285, 287
LERP, *see* linear interpolation
letters (symmetry of), 232
Levy C curve, 414
lexicon (definition of), 100
liar paradox, 330
licentious kitchen maid (painting), 669
lifeguard problem (minimization), 266
light (speed of), 337
likelihood vs. probability, 349
Lindemann, Carl Louis Ferdinand von (1852–1939), 8, 130, 135
Lindenmayer, Aristid (1925–1989), 418
Lindon, James Albert (1914–1979), 288
line
  vector equation, 452
line (parametric representation of), 679
linear interpolation, 477
  two parabolas, 448, 673
linear least squares, 435–439
linear perspective, 38, 510–546
Liouville constant, 129
Liouville, Joseph (1809–1882), 129
Littlewood's law, 351
Littlewood, John Edensor (1885–1977), 65, 351
Liu Hui (c. 225–c. 295), 140
Livio, Mario (1945–), 150, 151
lizard (Escher), 230
Lo-Shu method for a 9×9 magic square, 585
locus (definition of), 26
logarithmic spiral, *see* equiangular spiral
logarithms (definition of), 367
logical operations (on binary numbers), 49–50, 608
logistic map (chaotic behavior), 417
look and say sequence, 69, 75, 303
Loomis, Elisha Scott (1852–1940), 278
loop in a curve, 475, 503, *see also* cusp
Lorenz attractor, 417
Lorenz system, 417
Lorenz, Edward Norton (1917–2008), 356, 417
Lorenzo di Piero de' Medici (the Magnificent, 1449–1492), 515
lotus shape, 29
love curve, 42
love of God, 637
Loyd, Samuel (1841–1911), 553, 554
Lucas numbers, *see* Lucas sequence
  and powers of ϕ, 152, 296
Lucas primes, 296
Lucas sequence, 296–298, *see also* Lucas numbers
Lucas spiral, 296
Lucas, François Édouard Anatole (1842–1891), 296
Luther, Martin (1483–1546), 264
Lyapunov fractal, 420

M

M-Fibonacci (Mbonacci) sequences, 286
Máthé, Andras, 9
magic circles, 588
magic hexagon, 588
magic squares, 576, 583–589
magic theorem (wallpaper symmetries), 196–197
magic triangle, 588
magical magic square, 588
Mandelbrot set, 420
Mandelbrot, Benoît B. (1924–2010), 240, 357, 410, 415
  obituaries, 416
Manet, Édouard (1832–1883), 174
Manetti, Antonio di Tuccio (1423–1497), 513–517
mantissa (definitions of), 367
Marilyn vos Savant, *see* vos Savant, Marilyn
Mark Twain (1835–1910), 360
Marsili, Cesare (1592–1633), 262
Marx, Andrew, 9
Marx, Karl (1818–1883), 646
Mary Poppins (movie), 387, 388
Masaccio (Tommaso di ser Giovanni di Mone, 1401–1428), 517–520
Masolino da Panicale (Tommaso di Cristoforo Fini, 1383–1447), 518
mathematical beauty, xvi, 2, 4, 139
mathematical constants, 81
mathematical doodling, 43
mathematical equation (definition of), xvii
mathematical existence, xiii
mathematical jokes, xxiv
mathematical quotations, xxv–xxvii
mathematicians
  ages of, 560
  boring, 577
  curious, 35, 284, 309, 586, 662
  devious, 6, 379, 581
  funny and rigorous, 75
  hard workers and fun loving, 76
  human, 585
  methodical, 47
  meticulous, 287, 607
  recreational, 76
  square, 577
  with hats, 561
  young and original, 11
mathematics
  confusing, 152
  debates in, 272–273
  definition of, xxix
  emotional, xxi
  fun, 114
  I hate it, x
  is it science?, xix–xxiii
  nature of, 301
  oldest unsolved problem, 76
  origin of term, xii
  restrictions in, 267–272
  study of patterns, xiii, xxix
  unchanging, 6
  when will I use it?, xiv
  why so effective?, xvi–xviii, 104
  widely applicable, 421
Mathews, Daniel, 281
matrices
  diagonally dominant, 493
  nonsingular, 493
  orthogonal, 220, 221
  tridiagonal, 493
Maugham, William Somerset (1874–1965), x
Maxwell, James Clerk (1831–1879), xvi, 32, 352, 622
Mayfield Clarke, Marilyn (1931–1991), 341
meandering ratio, *see* sinuosity index
median (of a data set), 608
Medici, Giovanni di Lorenzo de (1475–1521), 264
Medici, Lorenzo di Piero de' (1449–1492), 264
Meisner, Gary B., 150
Melencolia magic square, 583
Menger sponge, 240, 415
Menger, Karl (1902–1985), 240
Merrill, James (1926–1995), 359
Mersenne, Marin (1588–1648), 262, 263, 331
Messing, Susan (1963–), 237
metadrome (definition of), 165
metallic ratios, 300–301
Meyer, Carl D., 463
mid-range average, 608
millillion, 66, *see also* large numbers (notation for)
miracle (mirrorless crossing), 194, *see also* glide reflection

Mittal, Adi, 154
mode (of a data set), 608
modulo 2 (same as XOR), 609
moiré patterns, 39–41
monotile, *see* Einstein tile
monster group, xxviii, 171
Monte Carlo method (for π), 136
Monty Hall problem, xx, 341–343
Morgan, Augustus de (1806–1871), xxi
Morgan, John Willard (1946–), 118
Moriarty, Liane (1966–), x, 352
Morse, Samuel Finley Breese (1791–1872), 359
Moser's circle sequence, 282–283, 667
Moser, Leo (1921–1970), 282
most irrational number (φ), 90, 150, 293, 295
MS Copilot in this book, 3, 8, 51, 60, 65, 95, 126, 252, 313, 318, 326, 329, 342, 350, 354, 388, 416, 559, 565, 567, 571, 576, 662, 664, 669, 672, 692
multiplicative magic squares, 587
multiplying negative square roots, 108, 142
Murata Constant, 81

N

Napier's number, *see* Euler's number
Napier, John (1550–1617), 131
Narayana's cows sequence, 295
naughty primes, 65
negate and exchange rule, 465, 657
negative square roots (multiplying), 108, 142
Nelson, Leonard (1882–1927), 343
Nepali flag, 618
Newcomb, Simon (1835–1909), 361, 362, 367
Newton polynomial, 480–481
Newton, Isaac (1642–1727), xii, xvii, 27, 263, 265, 308, 335, 480, 505, 520
Niels Fabian Helge von Koch (1870–1924), 414
nine-point circle, 549
(n, k) codes, 594
noble number, 86
noem (nth order extreme mean), 158, *see also* poem
noise, 440
  acoustic, 591
  as texture, 441
  atmospheric, 355
  correlated, 441
  decorrelated, 441
  electromagnetic interference, 591
  error-control codes, 440
  gradient, 441
  image of random pixels, 360
  in data transmissions, 591
  random, 440
  white, 440
non-collision singularities, 335
non-commensurate dimensions, 665
non-computable numbers, 71
non-palindromic primes, 663
nondifferentiable function, 427
nonlinear least squares, 439
nonlinear perspective, 510
nonlinear processes, 354, 356
  and chaos, 356, 357, 417
nonperiodic tiling, 222, *see also* aperiodic tilings
nonsense in this book, 668, *see also* jokes in this book
nonsingular matrix, 493
normal (definition of), 221
normal numbers, 71, 72, 637
nova fractal, 420
nowhere differentiable function, 427, *see also* everywhere discontinuous functions
number (definition of), 50
numbers, 47–166
  constructible, 71
  triangular, 54
  abundant, 71, 75, 283
  algebraic, 71, 256
  almost integer, 71
  amicable, 71, 283
  any base?, 284, 288, 663
  apocalypse, 76
  aspiring, 283
  base-φ, 288
  base-10, 49, 52, 60, 288
  base-16, 63
  base-2, 49, 52, 608
    length of, 53
  base-8, 58
  Belphegor, 65
  boring, 2–3, 68, 71
  brown, 71

colossally abundant, 75
complex, 103–114, 142–146
composite, 75
computable, 71
cyclic, 71
deficient, 71, 75, 76, 283
definable, 71
dungeon, 71
favorite, xxviii–xxix
Fibonacci, 71
figurate, 53
Goliath, 76
Granville, 71
harmonic, 309
heptagonal, 58
hexagonal, 57
highly abundant, 75
highly composite, 75
hypercomplex, 71, 112–115
hyperperfect, 71
hyperreal, 71
illegal, 71
irrational, 78–97, 639
  definition of, 78
kissing, 71
noble, 86
non-computable, 71
normal, 71, 72, 637
oblong, 71
octagonal, 58
pentagonal, 56
perfect, 57, 71, 76
polygonal, 71
practical, 71
primitive abundant, 75
pseudo-random, 356
quasiperfect, 71
quaternions, 71
rarely used, 71
rational, 78–97
  definition of, 78
real, 2–4, 98–103
really exist?, xiii
self-describing, 69
semiperfect, 76
sociable, 71, 283
spaghetti, 71
square, 53
superabundant, 75
  superperfect, 71
  surreal, 4, 71, 273
  transcendental, 71, 256
  trigonometric, 71
  twin, 71
  twin primes, 71
  vampire, 71
  witness, 71
numerical sequences, 281–303, 663

O

0 (orbifold translation indicator), 191, 195, 196
objective probability, 382
objects of constant width, 614–615
oblong numbers, 71
octagonal numbers, 58
octahedron symmetry, 653
octal numbers, *see* base-8
octanacci numbers, 287
octonions, 59, 113, 644
odd perfect numbers (unknown), 76
Ogawa, Yoko (1962–), 136
Ohm, Martin (1792–1872), 149
Oiler, *see* Euler
omniscience paradox, 258
operations on points, 469, 471
operations on vectors, 471, 472
orbifold (meaning of name), 189
orbifold notation, 171, 189–197
  examples, 197
  miracle, 194
order of a group element, 172
orrery (and irrational gear ratios), 94–96
orthogonal functions, 221
orthogonal matrix, 220, 221
orthogonal polynomials, 221
orthogonality (definition of), 221

P

p-series, 307
p-test, 307, 664
  proof of, 307
Pacioli, Luca Bartolomeo de (1445–1517), 149
Padovan sequence, 299
Padovan, Richard (1935–), 299

painter's paradox, 328
palindromes, 167, *see also* Universal Palindrome Day
  definition of, 233
  delayed, 234
  in dates, 234
  in music, 234
  in numbers, 234–235, 658
  in speech, 234
  in text, 233
  non-palindromic primes, 663
palindromic numbers, 234–235, 658
palindromic primes, 65
pandigital formula (definition of), 163
pandigital number (definition of), 163, 568
pandigital number (representation of), 72
Pappas, Theoni (1944–), 422
parabola, 26–27
  and Bézier curve, 509
parabolic blending, 481, 498–502
paradox
  Achilles and the fly, 249, 325
  antinomy, 318
  Aristotle's wheel, 330
  barber, 321
  Berry's, 342, 636
  Bertrand, 337–339
  Bertrand box, 339–340
  cable guy, 321
  Cantor, 258
  categories of, 317–319
  dartboard, 332–335
  definition of, 317
  dialetheia, 318
  Ellsberg, 383
  falsidical, 317
  Gabriel's horn, 327
  Galileo, 320
  horror, 324
  infinity in finite time, 335–337
  König's, 322
  liar, 330
  Monty Hall problem, 341
  omniscience, 258
  painter's, 328
  preface, 320
  Quinn, 321
  racecourse, 249, 319, 325
  rakehell, 322
  Richard's, 322
  Ross Littlewood, 249, 343
  self-deception, 324
  self-fulfilling, 324
  ship of Theseus, 318, 665
  sorites, 321, 665
  St Petersburg, 322
  Thomson's lamp, 249, 319, 326
  Tristram Shandy, 320
  unpleasant, 324
  veridical, 317
  Wang's, 321
  Zeno's, 240, 320, 325
paradoxes, 317–343
  coin rotation, 324
  potato, 328
  Sierpinski-Mazurkiewicz, 249, 250, 335
  Zeno, 261, 319–320
parallel projection, 537
  as a limit, 541
parametric cubic, 474–477, 483–489
  geometric representation of, 485
  Hermite, 483–489
  inflection points, 475
  noninteractive, 484
  PC, 475
parametric curves
  cubic, 474–477
  loops in, 475
  substitution of parameter, 510, 679, 681
parity (definition of), 593
Parker, Matthew Thomas (1980–), 72, 297, 585
parking (a greedy problem), 589–591
Pascal triangle, 504–506
Pascal, Blaise (1623–1662), xiii, 667
pathions, *see* trigintaduonions
pattern
  center of mass of, 195
  fundamental region, 193
  gyration symmetry, 194
  kaleidoscopic symmetry, 191
  miracle symmetry, 194
  translation symmetry, 195
Paul III (Alessandro Farnese 1468–1549), 264
Pauli, Wolfgang Ernst (1900–1958), 117
Pauling, Linus Carl (1901–1994), 390

Peacham, Henry (1546–1634), 233
Peano curve, 415
Peano, Giuseppe (1858–1932), 240
pedal line of a triangle, 549
Peirce, Benjamin (1809–1880), 134
Peitgen, Heinz-Otto (1945–), 416
Penrose tiles, 223–226
  and golden ratio, 226
Penrose, Roger (1931–), 151, 223, 226, 260
pentagon (interior angles of), 64
pentagonal numbers, 56
pentanacci numbers, 287
percents (fun with), 69
Perec, Georges (1936–1982), 586
perfect numbers, 57, 71, 76
periodic tiling, 222, *see also* aperiodic, nonperiodic, and semiregular tilings
Perlin noise, 440–450
Perlin, Kenneth H., 441
Perls, Friedrich Salomon, (Fritz, 1893–1970), 240
permutation group (definition of), 179
permutations
  and symmetry, 179
  cycle notation, 180
perpendicular (meaning of), 221
perspective, 38
  divergent, 527
  general case, 538
  linear, 38, 510–546
  nonlinear, 510
  shallow, 530, 537, 538
  special case, 538
  standard position, 538, 539
  vertical, 535
perspective projection
  Dürer's method, 527, 528
  mathematical treatment, 537–546
Perugino, Pietro (c.1446/1452–1523), 518
phi (ϕ), *see* golden ratio
philology (definition of), 138
phyllotaxis (and spirals), 291
pi (π), 135–142
  and 5, 56
  and pizza, 136
  approximations of, 140
  continued fraction of, 87, 88, 92, 97
  memorizing digits of, 138
  non-transcendental?, 142
  not very irrational, 89
  philology, 138
  rational approximations of, xxviii, 93
Pi Approximation Day, 136
Pi Day, 136
Pikhurko, Oleg, 9
pins (choosing 4-digit pins), 383–385
Pisano, *see* Leonardo of Pisa
Pisano period (Fibonacci sequence), 287
pitch
  in rotation, 455
pizza and π, 136
plaindrome (definition of), 165
Planck time, 249
plastic (meaning of term), 299
plastic ratio, 298–299
plastic rectangle, 299
plastic spiral, 299
Plato (428/427 or 424/423–348/347 B.C.), 73
Platonic solids, 56, 169, 377
  as edge-transitive, 378
  the simplest, 184
Plutarch (c. A.D. 46–after A.D. 119), 318
Poe, Edgar Allan (1809–1849), 97
poem (pth order extreme mean), 158, *see also* noem
Poincaré, Jules Henri (1854–1912), 414
points
  control, 503, 508
    and convex hull, 508
    reversing, 508
  inflection, 475
  operations on, 469, 471
points and complex numbers, 113, 469
points and vectors, 469–472
Poirot, Hercule (fictional detective), 343
Poliziano, Angelo (1454–1494), 527
polygon (definition of), 135, 171
polygonal numbers, 71
polyhedra
  fair by continuity, 379
  fair by symmetry, 378
  Platonic solids, 377
polyhedron (definition of), 171
polyline, as a curve with maximum tension, 497
polynomials

Bernstein, 507
cubic, 474–477
definition of, 474
orthogonal, 221
wiggle, 475, 477
polytope
definition of, 171
the monster, 171
popcorn function, *see* Thomae function
Portrait of Dr. Gachet (book), 529
potato paradox, 328
Poundstone, William, 140, 319
power sets
cardinality of, 257–258
definition of, 257
practical numbers, 71
preface paradox, 320
Prelog, Vladimir (1906–1998), 455
prime constant, 75
prime number theorem, 73
prime numbers, 73–74
beastly, 65
definition of, 73
irrational square roots, 81
left-truncatable, 74
naughty, 65
not squares, 81
palindromic, 65
right-truncatable, 74, 637
primitive abundant number, 75
Prince of Pi, *see* Shaw, Lawrence N.
probabilistic counting, 403–407
probabilities in a continuum, 332–335
probability, 345–408
and likelihood, 349
and psychology, 383–385
and statistics, 348
birthday paradox, 385–386
choosing a candidate, 387–389
examples, 390–401
finding a birthmate, 386
in Vernam cipher, 689
objective, 382
of sum of events, 348
range of, 348
subjective, 382–383, 669
probability concepts, 345–347
probability density
definition of, 333, 334
nonuniform, 334
Wikipedia definition, 335
product of two negatives is positive, 272
pseudo-random numbers, 356
pseudoperfect number, *see* semiperfect number
psychology (and probability), 383–385
pursuit curve, 28
Putnam, Hilary Whitehall (1926–2016), xviii
puzzles, 551–579
14-15 puzzle, 553
16-room puzzle, 553
clock hands overlap, 562
coin puzzle, 555
dissecting a square, 552
Euler Trail, 555
vanishing area puzzle, 554
Pythagoras (c. 569–475 B.C.), 50, 149, 151, 178, 277
Pythagorean means, 607
relation of, 607
Pythagorean theorem, 277–279

Q

quadratic Bézier curve, 509
quadratic convergence (in AGM), 137
quadratic mean, 607
quadratic polynomial (a plane curve), 475
qualitative data, 363
quantitative data, 363
quartic polynomial
and ϕ, 154
definition of, 154
quasi-numerical sequences, 303–304
quasiperfect numbers, 71
quaternions, 71, 113, 115–128, *see also* hypercomplex numbers
and Euler equation, 123–127
and rotations, 450
and spin matrices, 117
quatrain (definition of), 235
Quine (self-replicating software), 12
Quine, Willard Van Orman (1908–2000), 12, 317, 318
Quinn's paradox, 321
quintic equations (no solutions), 277

R

racecourse paradox, 249, 319, 325
rakehell (meaning of), 322
rakehell paradox, 322
Ramanujan, Srinivasa Aiyangar (1887–1920), xxviii, 65, 74
  magic square, 588
Ramsey theory, 13–15
Ramsey, Arthur Michael (1904–1988), 13
Ramsey, Frank (1903–1930), 13
random noise, 440
randomness, 352–360
  and Benford's law, 362
  and information, 358–360
  in a die, 376–380
Rao, Michael, 227
rarely used numbers, 71
rational numbers, 78–97
  definition of, 78
  dense, 79, 255
real harmonic series, 309–312
real numbers, 2–4, 98–103
  continuum, 4, 99
  smallest, 4
Record, Taylor I. (1846–1912), 645
Recorde, Robert (c. 1510–1558), 49
redundancy
  definition of, 592
  in a natural language, 593
  in digital data, 592
Reed, Irving Stoy (1923–2012), 598
Reed–Solomon codes, 598–604
reflection matrix, 216
reflections
  in a cube, 182
  in a triangle, 179
  in three dimensions, 181
  in two dimensions, 181, 652
  rotary, 182, 183
regular tiling, 222, *see also* aperiodic, nonperiodic, and semiregular tilings
relativity, special, 320
Rényi, Alfréd (1921–1970), 24
repeating 9's, *see* 0.9
repetend (definition of), 78
repfigit number, *see* Keith number
Reuleaux triangle, 614
Reuleaux, Franz (1829–1905), 614
reverspective (reverse perspective, 527
Rhind, Alexander Henry (1833–1863), 645
rhombic dodecahedron, 378
rhombic triacontahedron, 378
Rich, John, 258
Richard's paradox, 322
Richard, Jules (1862–1956), 322
Richter-Gebert, Jürgen, 197
Riemann series theorem, 305–307
Riemann zeta function, 307, 316, 665
Riemann, Georg Friedrich Bernhard (1826–1866), 105, 429
right-hand rule, 673
right-hand rule (in rotation), 453
right-truncatable primes, 74, 637
Roberval, Gilles Personne de (1602–1675), 263
Robinson, Julia Hall Bowman (1919–1985), 78
Rodrigues' rotation formula, 123, 460–461
Rodrigues, Benjamin Olinde (1795–1851), 461
Roestraten, Pieter Gerritsz van (1630–1700), 511, 669
Rogers, William Penn Adair (1879–1935), 360
Rohmer, Éric (1920–2010), 332
roll, 455
Roman numerals, 663
rook's tour, 576, 577
Rooney, Andrew Aitken (1919–2011), 408
root mean square, *see* quadratic mean
roots of $i = \sqrt{-1}$, 111, 142, 144, 641
roots of unity, 107–108, 641
  and product of chords, 615
Rorschach inkblot test, 54
Rosenberg, Joel (1954–), 617
Ross-Littlewood paradox, 249, 343
Rostand, Edmond Eugène Alexis (1868–1918), (End of index)
rotary reflections, *see* turn reflections
rotary reflections of the cube, 182, 183
rotation
  CORDIC, 464–469
  equals two reflections, 218
  extrinsic, 455
  followed by a translation, 219
  Givens, 461–462
  intrinsic, 455

of complex numbers, 250
pitch, 455
Rodrigues' rotation formula, 123, 460–461
roll, 455
three-dimensional, 450–469
yaw, 455
rotation group, 450
rotation matrix, 216
rotoreflection, 183
an improper rotation, 187
in a tetrahedron, 185
rounding a random fraction, 99, 396, 397
RS codes, *see* Reed–Solomon codes
ruled surfaces, 41
Runge's phenomenon, 475
Russell, Bertrand Arthur William (1872–1970), 10, 11, 321, 360, 394

S

Sabey's formula, 163
weakly pandigital, 651
Sabey, Richard, 163
Sagan, Carl Edward (1934–1996), 140
Sagrada Familia magic square, 583
Saltzman, Cynthia, 529
Santa Maria del Fiore (Cathedral), 513
Sarközy, Nicolas Paul Stéphane de Nagy-Bocsa (1955–), 416
Sautoy, Marcus du (1965–), 187
Schaaf, William Leonard (1898–1992), 142
Schiaparelli, Giovanni Virginio, (1835–1910), xviii
Schilpp, Paul Arthur (1897–1993), 410
science and superstition, xx
Scott, Walter (1771–1832), 235
sedenions, 113, 644
self similar
ϕ?, 160
golden rectangle, 160
golden spiral, 160
self-deception paradox, 324
self-describing numbers, 69
self-fulfilling belief paradox, 324
self-inverse fractals, 414
self-reference (definition of), 12
semimagic square, 586
semiperfect number, 76
semiregular tilings, 223
sequences, 280–304
2, 3, 4, 82,000, 284
aliquot, 64, 283–284
comma, 301–302
extended Fibonacci, 161, 296
Fibonacci, 285–296
hyperoperation, 47–49
Lucas, 296–298
Moser circle, 282–283
numerical, 281–303
quasi-numerical, 303–304
Stern-Brocot, 254
Sergels torg (a superellipse), 37
series, 304–316
convergence, 304, 307
harmonic, 54, 309–314, 432
diverges, 51
real, 309–312
p-series, 307
singular and plural, 304
Serle, Rebecca, 346
set theory, 251–257
Seurat, Georges-Pierre (1859–1891), 264
seven deadly sins, 58
seven-segment display, 58
Shakespeare (and letter distribution), 358
Shakespeare, William (1564–1616), xxix
Shannon, Claude Elwood (1916–2001), 592, 635
Shaw, Lawrence N. (Prince of Pi, 1939–2017), 136
shearing matrix, 216
Ship of Theseus paradox, 318, 665
show and tell sequence, xxviii, 69
Sierpiński's constant, 81
Sierpiński, Wacław Franciszek (1882–1969), 240, 414, 422, 424
Sierpinski carpet, 240, 414, 422
Sierpinski triangle, 412, 414, 415, 422
the fractal dimensions, 412
Sierpinski-Mazurkiewicz paradox, 249, 250, 335
sign function, signum, 427
significand, *see* mantissa
silent letters (not V), 358
silver ratio, 300, *see also* metallic ratios
Simmons, Dan Hermann (1938–), 631
Simson line of a triangle, 549

singleton (one-element set), 257
singular value decomposition, *see* SVD image compression
sinh ("shine," definition of), 144
sinuosity index (of a river), 141
Sloane, Neil James Alexander (1939–), 281
slope, *see also* tangent vector
  definition of, 474
smallest positive real number, 4
so, what is chance? (quotation), 357
sociable numbers, 71, 283
Solomon, Gustave (1930–1996), 598
somewhat-magic square, 587
sorites paradox, 321, 665
space filling curves, 235
spaghetti numbers, 71
sphere decompositions, 11
spirals, 28–29
  logarithmic, 160
spline (definition of), 481
spline interpolation, 481–503
splines
  as a piecewise curve, 475
  cardinal, 481, 493, 495–498
  cubic, 481, 490–494
    clamped, 493
    Hermite, 483–489
  tension in, 493, 495–498
square numbers, 53
square spiral, 28
squaring the circle (impossible), *see* dissecting the circle, 8–9, 135
St Petersburg paradox, 322
statistics (and probability), 348
Stern-Brocot sequence, 254
Sterne, Laurence (1713–1768), 320
Stigler's Law, 361
Stigler, Stephen Mack (1941–), 361
strange attractor
  definition of, 416
  examples of, 416
strange attractors (fractals), 416–418
strong law of small numbers, 282, 667
subjective probability, 382–383, 669
submarine (hunting), 557–558, 686, 687
substitution of parameter (in curves), 510, 679, 681
sunflowers and the irrationality of ϕ, 292–295
superabundant number, 75
superegg, 37
superellipse, 35–37
supergolden ratio, 295
supergolden rectangle, 296
superperfect numbers, 71
Suppes, Patrick Colonel (1922–2014), 347
surjective function (definition of), 173, 433
surreal numbers, 4, 71, 273
SVD image compression, 604–606
syllables (in the names of integers), 636
symmetric group
  S_4, 185
  definition of, 180
  not the same as symmetry group, 180
symmetricity, 168
symmetries
  of a brick wall, 187, 653
  of tetrahedron, 183–186
symmetry, 167–237
  ambigram, 233
  and permutations, 179
  bilateral, 168
  chirality, 168
  fundamental region of a pattern, 193
  gyration, 194
  historical developments, 169
  in a circle, 169
  in dice, 180
  in graphics, 235
  in poems, 235
  in text, 233–234
  in three dimensions, 180
  in two dimensions, 181
  inverse of, 175
  isometry, 171
  kaleidoscopic, 191
  miracle, 194
  of a cube, 182
  of the rhombus, 177
  of the square, 178
  of the triangle, 178
  palindrome, 233–235
  poem on, 237
  reflection, 216
  rotation, 216
  scaling, 216
  shearing, 216

translation, 195
trivial group, 176
symmetry groups, xxviii, 170–186, 378
definition of, 175
elements of, 175
operation of, 175
transitive, 378
trivial, 176
Syrus, Publilius, 604

T

T-square (fractal), 415
Tait-Bryan angles, 455
Talwalkar, Presh, 136
Taneja, Inder Jeet, 72
tangent vector, 481, 488
Bézier, 508
definition of, 473, 474
direction of, 676
extreme, 491, 493
indeterminate, 489, 683, 684
of a PC, 476, 477
Tangram, 9
and circle dissection, 9
Tarski, Alfred (1901–1983), 9
Tartaglia, Niccolò Fontana (1499/1500–1557), 104
Tegmark, Max (1967–), xviii
teleidoscope, 188
ten commandments, 53, 60
Tennyson, Alfred (1809–1892), 236
tension (in curves), 493, 495–498
tessellation, *see also* tiles and symmetry
definition of, 222, 228
how to, 228–230, 659
same as tiling, 228
tesseract (4-dimensional cube), 171
tetrahedron
no center, 184
symmetries of, 183–186
tetranacci numbers, 287
tetration, 145, 251
definition of, 146
of i, 144, 146–147
repeated exponentiation, 133
Text by ChatGPT, vii, xii, xxix, 76, 129, 130, 162, 166, 170, 237, 274, 352, 408, 411, 422, 578, 590, 652, 665, 671
texture
definition of, 441
procedural, 441
The Prisoner, television series, 608
thinking outside the box, 554
Thomae function, 432, *see also* popcorn function
Thomae, Carl Johannes (1840–1921), 432
Thompson, D'Arcy Wentworth (1860–1948), 150
Thomson's lamp paradox, 249, 319, 326
Thomson, William (1892), 119
three-dimensional rotations, 450–469
4 numbers needed, 118
Thurston, William Paul (1946–2012), 118, 189, 190
tic-tac-toe board, 15, 16
tiles and symmetry, 222–230, *see also* tessellation
tiling
aperiodic, 222–228
Einstein, 226–227
nonperiodic, 222
periodic, 222
regular, 222
semiregular, 223
Wang, 227–228
time (is money), 323
Toblerone chocolate as a polyhedron, 379
Tolkien, John Ronald Reuel (1892–1973), 171, 537
Tolstoy, Lev Nikolayevich (1828–1910), 62
Torricelli's trumpet, *see* Gabriel's horn
Torricelli, Evangelista (1608–1647), 262, 327
Trahan, Alexander C., 374
transcendental numbers, 71, 98, 129–130, 256
i^i, 144
difficult to prove, 135
no compass and ruler, 8
rare, 98
uncountable, 98
triacontahedron, *see* rhombic triacontahedron
triangle
pedal line of, 549
properties of, 54
Simson line of, 549
triangular numbers, 54

increasing monotonic, 75
tribonacci numbers, 287
tridecimal numbers, *see* base-13
tridiagonal matrix, 493
trigintaduonions, 643
trigonometric numbers, 71
trimagic square, 586
trinion, *see* hypercomplex numbers
trinions (3D numbers), 112–113, 115
Trinity (fresco), 517
trinomial theorem, 506
triple product, 453
Tristram Shandy paradox, 320
trivial group, 176
as a subgroup, 652
Tron (movie), 441
truly large numbers (law of), 350–351
Tupper's self-referential formula, 12
Tupper, Jeffrey Allen, 12
Turgenev, Ivan Sergeyevich (1818–1883), 484
twin numbers, 71
twin primes, 71
two-dimensional transformations, 198–221, 425
types of integers, 75
Tyson, Neil deGrasse (1958–), 667

U

Uccello, Paolo (1397–1475), 516
Ulam, Stanislaw Marcin (1909–1984), 240
uncertainty principle (and π), 141
unconditional convergence (of series), 305
Unger, Spencer, 9
uniform convergence (of series), 305
uniform tiling, *see* semiregular tilings
unity
fifth roots of, 109–110
roots of, 107–108, 641
Universal Palindrome Day, 234
Universal parabolic constant, 81
unpleasant paradox, 324
up-arrow notation, xxix, 48–49, 66–68

V

Vázsonyi, Andrew (1916–2003), 5
Vail, Alfred Lewis (1807–1859), 359
vampire numbers, 71
Van der Waerden's theorem, 14
vanishing area puzzle, 554
vanishing points, 530
examples of, 531
rule of, 530, 538
Vasari, Giorgio (1511–1574), 514
vectors
addition, 472
cross product, 451–453, 673
direction of, 452, 673
cross-product, 472
dot product, 221, 472
no definite location, 457
operations on, 471, 472
orthogonal, 472
triple product, 453
vector product, 451
velocity of curves, 477
veridical paradox, 317
Vernam cipher, 611, 689
Vernam decoding, 611
Vernam, Gilbert Sandford (1890–1960), 611
Verne, Jules Gabriel (1828–1905), 62
versor (a unit quaternion), 121
vertical perspective, 535
visible math objects (graphics), 19–41
Vitali's set, 11
Viviani's theorem, 400
Viviani, Vincenzo (1622–1703), 262
Vogel's model of sunflower spirals, 293, 295
Vogel, Helmut (1929–1997), 295
vos Savant, Marilyn (1946–), 341

W

Wagon, Stanley, 577
Waldo, Clarence Abiathar (1852–1926), 646
Walker, Paul Robert (1953–), 18, 518
Wallis Constant, 81
Wallis, John (1616–1703), 88, 105, 263, 265
wallpaper patterns, 190, 195
Wang tile, 227–228
Wang's paradox, 321
Wang, Hao (1921–1995), 227
Wantzel, Pierre Laurent (1814–1848), 562
Weierstrass function, 414, 427–429
Weierstrass, Karl Theodor Wilhelm (1815–1897), 266, 414, 427, 428, 430
weird number, 76

Wells, Herbert George (1866–1946), 398
Wessel, Caspar (1745–1818), 105
West, Mary Jane (Mae, 1893–1980), 474
when will I use all this math?, xiv
white noise, 440
Wieferich primes, 65
Wigner, Eugene Paul (1902–1995), xvii
Wiles, Andrew John (1953–), xxii
Wilson, Sloan (1920–2003), 664
witch of Agnesi (curve), 35
witness numbers, 71
word size in English, 689

X

Xia, Zhihong "Jeff" (1962–), 336
XOR, *see* exclusive-OR

Y

Yang, Chen-Ning (Frank, 1922–), 16
yaw, 455
Yogi Berra, *see* Berra, Lawrence Peter

Z

Zapf, Hermann (1918–2015), 166
Zeckendorf's theorem, 288
Zeckendorf, Edouard (1901–1983), 288
Zedong, Mao (1893–1976), 232
Zeno
  paradoxes, 240, 261, 319–320, 325
  racecourse paradox, 319
Zeno of Elea (c. 490–425 B.C.), 240, 261, 314, 319, 320, 325
Zermelo, Ernst Friedrich Ferdinand (1871–1953), 10
  origin of name, 10
zero, 51–52
  a place holder, 51
  and the calendar, 51
  dividing by, 52
  refuge of the divine spirit, 52
zigzag (scanning pattern), 245, 253, 557
Zu Chongzhi (429–500), 140
Zuse , Konrad Ernst Otto (1910–1995), 635

A great nose may be an index of a great soul.
—Edmond Rostand, *Cyrano de Bergerac, (1897)*.

Colophon

This book was started in early 2007 and has continued, after a long break, in mid 2023. The book was designed and typeset by me with the TEX typesetting system developed by Donald Knuth. The text and the tables were done with Texlive on a Macintosh computer. The drawings were generated by me in Adobe Illustrator, also on the Macintosh. Figures requiring calculations were computed first in Mathematica, GeoGebra, Desmos, or Matlab, and then "polished" in Illustrator. A few poems and about 25 images (one of them on this page) were generated by chatGPT, versions 4o and 5 and by Microsoft Copilot. Together with the many jokes sprinkled throughout the text, they were used to fancify and embellish the sometimes dry mathematical content.



The TEX Lion. Image by MS Copilot, Sept. 2025.

Some said, "David, publish it;" others said "Not so."
Some said, "It might do good;" others said, 'No."
—John Bunyan (paraphrased).