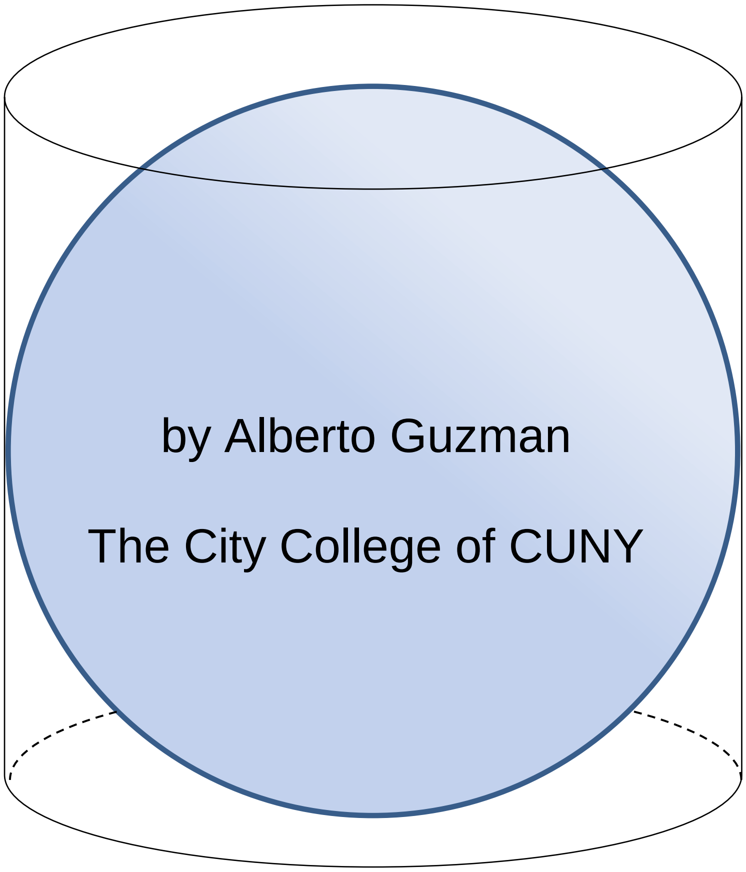
Mathematics in Historical Context

The Development of

Geometry, Algebra, and Number Theory



by Alberto Guzman

The City College of CUNY

Preface

This material attempts to put the development of geometry, algebra, and number theory into historical context. It cannot hope to be comprehensive, and therefore makes frequent references to classics that do give comprehensive coverage of the history of mathematics. It does not try to be linear: It treats topics for their interest, even if they are not links in a chain of development, and sometimes puts them out of chronological order. It is heavier on the math than on the history, reflecting thereby the author's predilection.

I will refer to this document as the "book." The usage reflects my initial fantasy that it would be published to great acclaim. At some point, however, I decided that there are conventions of book writing I don't want to adhere to, including proscriptions against speaking in the first person, using contractions, and ending clauses with prepositions. For that reason, the book will reside online. I figure that this medium will allow me freedom to express myself in my peculiar style, some of which may now be obvious to you. It will further allow me to keep the discussion conversational, very much like a class.

There is one convention on mathematical style that I will follow: A word will appear in **boldface** to indicate that it is being defined, or that its definition is close by. To *stress* a word, I will put it in italics.

> You will find some paragraphs indented and bordered on the left, like this one. Those are the places where the algebra, arithmetic, or geometric argument gets detailed. [Also, you will find asides, editorials, and other tangential items that I think will interest you, set off in square brackets.]

The decision to put the book online carries advantages for the reader, as well. The first is that the book is free. Second, it is widely available, especially now that mobile devices can access it from what seems like everywhere. More fundamentally, it now includes hyperlinks to outside sources and hyperlinked cross-references among different parts of the book.

If your device can download the book, you should do so. Accessing it every time on City College's website may be a slow process. In addition, if you access it there and click an external link, you may lose your place in the book. [It depends on your browser. If you are viewing this via the College and want to check on what I mean, click here, then hit the "Back" button.]

None of the historical material here is original. Neither is the mathematics, although often the sequencing, treatment, or way of looking at topics is my idea. (I am happy to say that I managed to concoct many of the proofs without help.) I learned some things from books. When I can, I identify the source. Others I learned from my teachers. With few exceptions, I cannot specify who they were, even though I am proud to have been their student.

I am not entirely confident in my knowledge of history. Accordingly, I make this offer: If you find an error or inaccuracy in the historical narrative and can provide correction, together with a published reference or reliable internet link arguing for the correction, then I will insert it into the book and name you as its source. For that matter, if you find a good reference in support of what I write, I will be happy to include it. I am confident about my mathematical knowledge, but will make a similar offer there, asking you to provide not a reference but a mathematical argument by way of justification.

The material was put together for a History of Mathematics course taken by Master's Degree candidates in Mathematics Education. For that reason, it sometimes addresses people who will teach high-school or middle-school math. Since it was not originally for math majors, it tries to confine itself to the mathematical level of high-school geometry, intermediate algebra, and trigonometry. I think it succeeds through the first six chapters. In that part, some material and (especially) exercises call for, or at least reward, knowledge of the calculus. Indication to that effect is given at those places. Of course, eventually you have to get to the development of the calculus. There, you will find discussion of calculus-related (“analytical”) concepts in geometric and algebraic terms. The discussion avoids presupposing undergraduate training in the topics at hand.

Notwithstanding the elementary nature of about half the material, the treatment displays and demands of the reader some mathematical sophistication. Mostly that means a feel for the nature of proof. The demand increases in the last third, as the underlying math becomes abstract. In the first two sections of Chapter IX, the thinking [even without requiring calculus] rises to the advanced undergraduate level.

The format resembles an outline because it was originally put together as a course outline, without the intervening paragraphs of text.

Alberto Guzman

October 2010 to around June 2019

I wasn’t in a hurry.

Those who can, do;
those who can’t, teach.
GB Shaw

The ideal condition would be, I admit,
that men should be right by instinct;
but since we are all likely to go astray, the reasonable thing is
to learn from those who can teach.
Sophocles

My money’s on the Greek.
A Guzman

Table of Contents

Chapter and section listings are links. Click on a title to get to that part.

Chapter I. Prehistoric Peoples

It is a signal event when we find evidence of how people lived and what they did before the advent of writing. Our evidence is mostly fragmentary: pieces of pottery, tools, weapons. The famous Lascaux paintings are an incredible large-scale souvenir. Still, there are types of mathematical thinking we can reasonably ascribe to early man. This chapter consists of such informed guessing.

Section I.A. Hunters

Early humans were nomadic hunters. Their lives would have been a constant struggle for survival. Their thinking would have been devoted to securing food and protecting the family. Life would not have allowed opportunity for a leisurely intellectual pursuit. What kind of mathematical thinking could they have done?

1. Geometry

Any creature needing to hunt for a living (including picking from the ground and out of trees) has to process some geometrical information. It must be able to judge and think in terms of spatial relationships: near and far, level and inclined, clear and wooded. If it has the power to communicate, it can profit from splitting a hunting party into a part going *this side* of the river, a part on the *other side*. It can worry that the kids, failing to listen, will go *beyond* the hill. (See Exercise 1.)

One particular concept that must have resided in early minds is *symmetry*. Designs painted onto surviving prehistoric pottery reveal awareness of symmetric patterns. Separately, observe that many land creatures are left-right symmetric. (See Exercise 4.) Recognizing symmetry therefore helps distinguish between living forms, to be sought as prey or avoided as predators, and features of the landscape, like plants and rocks.

2. Arithmetic

Some forms of arithmetic (arithMETic, the adjective) thinking would have been accessible and profitable to primitive humans.

At the most basic level is counting. If you can count, then for example you can tell whether everybody from your group is present without having to match people up with a mental list of names or faces. You can tell whether part of the group has been gone for an unusual number of days. You can think in terms of *how many* of these things, say fruit, you might trade for how many of those, say ropes.

With counting, some sophistication is reachable. You can handily (no pun) count up to 10 on your fingers, and up to a few dozens with nuts or stones or such things. Counting big numbers, like the number of days between the first snowfall of this year and the first of next year, must have required use of a long stick or bone on which to make marks. This last suggests that people must quickly have adopted **aggregates**. You know how we mark

|, ||, |||, ||||, 𝍸,

so that

𝍸 𝍸 𝍸 |||

is easily recognizable as 3 × 5 + 3 without counting 18 marks. If you let special marks, like deep or extra-long notches or especially big stones, represent fives or tens or a different meaningful aggregate, then you expand your counting reach. You can even do that with your fingers. If you let the fingers on one hand represent fives, then counting fives on that hand and ones on the other allows you to count to 30 using just your hands.

At a higher level lie the operations. We naturally think of addition first, because it is what we learn first. But that sequencing is based on pedagogical considerations: Addition is conceptually easy. It might have been that early humans did something like division first, based on the need to separate a group of people or objects into equal or nearly equal subgroups. In any case, the ability to add would have become valuable, since it lets you decide whether putting together two collections having the same purpose renders the combined collection sufficient for some function neither one could serve alone.

Algebra—in the sense of creating descriptions (for us, equations) from which information may be drawn—sounds like an area of thought unlikely to have occupied early humans. It is difficult to imagine ways such thinking could have been advantageous to hunter-gatherers. Indeed, we will see later that evidence of "algebraic thought" suggests that its origin had to wait for complex social structures.

3. Astronomy

The subject is not strictly mathematical, but astronomy played an important role in the development of mathematics. Early men would have profited from astronomical knowledge. Hunters gazing upon the night sky must have observed that the stars, whatever they might be, reside in patterns whose positions change in tune with the seasons. For example the constellation Scorpius (which actually does resemble a scorpion) is (in our era and location) as high as it gets in the South at sunset in early September. Over the next two months, its sunset position drifts westward. This seeming westward drift is due to Earth's revolving about the Sun. As we orbit the Sun, it appears to us that the Sun is drifting *eastward* against the starry background. The drift causes the Scorpion to approach the sunset, so that before December the constellation becomes invisible in the twilight. In its prominent September position, Scorpius announces (for the northern temperate zone) the arrival of autumn. Humans aware of that signal would have been warned of the impending migrations of the animals, the changing of the trees, and the need to prepare for the season, three months away, of cold and darkness and death.

Even more noticeable than the harmony between stars and seasons would have been the rhythm of the Moon. The Moon is visible most of the time, and it is reasonably regular. Its cycle, full Moon to full Moon, averages about 29½ days. (In truth, the actual length varies by hours over the course of multiple "lunations", but the variation is difficult to spot without a timepiece.) The length, hereafter "the moon," is not some modern scientific determination. One human could calculate the average by watching 100 lunations. (How long would that take?) Suppose he has [like me] trouble telling the night of full Moon from the one or even two nights before or after. Counting the number of days between one full Moon and the one 100 cycles later would give him the average moon to within 4/100 days, or about 1 hour. For that matter, doing the same from one *lunar eclipse* to another 100 or so cycles later would give the moon to within hundredths of an hour. The regularity of the Moon allows its use as a device to time reasonable numbers of days, and even as a somewhat clumsy device to time the seasons. (Very clumsy: The number of days in the cycle of the seasons, hereafter "the year," falls between multiples of 29½).

Exercises I.A.3

1. What other kinds of geometric or arithmetic thinking would hunter-gatherers have done?
2. Do animals (other than man) do anything like geometric reasoning? like arithmetic reasoning?
3. Why would it have been important for hunters to think in terms of *near* and *far*? Would their thinking have been quantitative, using measurements ("*such and such is 50 rods away*")?
4. a) Many terrestrial animals, humans among them, are (left-right) symmetric about a vertical plane perpendicular to their fronts, but not about any vertical plane parallel to their fronts, nor about any horizontal plane. Why is that?

b) What advantage accrues to an animal whose structure is symmetric?
c) Is asymmetry ever advantageous?

5. We are highly conscious of minutes, hours, days, weeks, months, and years. Which of these units of time would early man have been aware of and considered important?

Section I.B. Farmers

Somewhere between 10,000 and 20,000 years ago (estimates vary), humans realized that they could *grow* food. The discovery of agriculture ended man's reliance on the unreliable and risky process of hunting, and substituted a source of nourishment that is at once more dependable and more amenable to storage. With people farming in one place, at least as long as the land was fertile, new areas of mathematical thought would have opened up.

1. Geometry

People staying in place would have built permanent shelter. Therefore they would have begun to learn principles of architecture. Those include, aside from the properties of materials, much geometric thinking. For example, people must have learned early on that doubling the dimensions of a surface, like a roof, quadruples its weight. In making supports, they must have learned that triangles are rigid, as say quadrilaterals are not. Form three sections of a folding carpenter's meter into a triangle and hold, between one thumb and forefinger, the start of the first section against the end of the third. Pulling on the sides will show you that as long as the joints and the sides do not break, the triangle retains its shape. Then form four sections into a square and hold beginning to end. A pull on the corner diagonally opposite your hold deforms the square into a diamond. The triangular bracing we see on exposed steel skeletons, as on a bridge, must be one of the earliest geometric ideas put into architectural use.

A second pursuit that would have rewarded geometric thinking is what we will begin to call **surveying**. The word "survey" has modern French origin, but the craft of land measurement is ancient. "Straight" is related to "stretch" and "line" to "linen," and therefore to rope. Farmers would have fashioned and used ropes to lay out plots of land, set boundaries, and of course measure. The geometric principle that a circle surrounds the greatest area for a given length of boundary—or for rectangles, that the square has the same property—must have been a commonplace to experienced "rope stretchers."

One final area of geometric thought, for which we do have pieces of evidence, is art. Pieces of pottery that have come down to us demonstrate beyond doubt awareness of, among other concepts, congruence, similarity, symmetry, and recurring patterns that fill up an area ("tessellations").

2. Arithmetic

As far as numeration, there is no obvious need farmers would have that their hunting forebears would not already have met. If you produce a quantity of grain, vegetables, or fruit, you would not normally have a reason to count them. You would, however, have reason to divide one quantity into multiple shares. That would have you thinking about fractions and about ways to compare them or to operate on them as you do with whole numbers.

3. Astronomy

We have already noted the importance of the stars as harbingers of the seasons. Their role is muted near the Equator. In much of the tropics, there are two seasons: either the times when the Sun is dead overhead (March and September) versus its northern and southern extremes (June and December); or else the rainy season versus the dry. In the temperate zones, though, the stars mark the times of planting, harvesting, and the other activities essential to agriculture.

Section I.C. Civilization

At some point, people realized the advantages of cooperative production. Agreeing to grow large crops and divide the yield, or to grow a variety of crops and trade yields—not to mention to act in concert to defend against raiders—would have led to increasingly large farming communities. These would over time begin to merge, eventually (at least 8000 years ago) into cities. In such places, the majority of residents would be far from the sources of food. They would have acquired “jobs,” activities that produce goods or labor they can trade for necessities. [We could choose to *define* “civilization” as a plan of human organization in which people can fill their needs by tending to the needs of others.]

Civilizations arose along four river valleys. (Name them.) Those places get flooded every year. (Why would people choose to live on land subject to flood?) Trying to make a life there demands harnessing the floods. Accordingly, people drained swamps, built levees and dams to control the flows, dug canals and other channels to provide more control plus irrigation and storage. All this civil engineering required large-scale central planning. The growing communities sprouted administrative centers, leading to governmental arrangements. What math would they have needed?

1. Geometry

The needs of surveyors and architects immediately leap to mind. Planning and laying out irrigation and drainage networks demand skill in measuring distances and angles and in dealing with similarity. Perhaps greater demands fell to architects, as the increasing concentration of people led to buildings of administration and governance. These began to be built at big and massive scale, to convey the importance of the places, and of course of the occupants. People also began to build ceremonial centers, in particular religious places. The temples that have survived indicate deep knowledge of geometric concepts like symmetry and properties of regular figures.

The rise of religion implies the creation of groups liberated from physical labor, having therefore “leisure” time in which to acquire and develop specialized knowledge. This “priestly class” could have stored knowledge without regard to its application to the community’s life.

2. Arithmetic

Because the rise of cities demanded central administration, the administrators needed to master **logistics**, the art or craft of managing the flow of people, tools, goods, whatever goes into doing large-scale projects. An immense number of things needed planning: managing the flood plain; how, where, and in what quantities to plant crops; how to create paths for the movement of humans and supplies; levying and allocating taxes to raise the payments the administration has to make; assembling and maintaining an army to defend the community. These things demanded, separate from some right-brained (geometric) thinking, enormous capacity to calculate. Whatever their numbering system was, people must have valued computational skill, and no doubt computing devices.

3. Astronomy

The chief feature of astronomy in the coming of civilization must be that it became a viable job. If your knowledge of the sky lets you make predictions, especially of spectacular events—the coming of the floods and eclipses of the moon are two obvious candidates—then you can easily sell yourself as a reader of the future. Recall that we referred to a priestly class, having time for such activity as observing the heavens. They would have jealously guarded whatever knowledge they derived. It would have been valuable, enhancing their standing with the addition of “oracle” to their titles.

Chapter II. Early Historical Peoples

Now we look at the mathematics of two civilizations from the early historical era, the time after the invention of writing. [That's about 5000 years ago. In your mind, distinguish between pre-historic and pre-civilized. The latter, as we said, goes back 8000+ years.] The development of writing is an epic event in the life of our species. When, scores of thousands of years ago, humans evolved the ability to talk, we acquired the ability to profit from the knowledge and experience of others. With the advent of writing, it became possible for us to access the knowledge of people removed from us in time as well as place. [It became possible for teenagers to ignore predecessors beyond their parents.]

Section II.A. The Egyptians

The valley of the Nile was protected by desert, mountains, and distance from the sea. That stands in contrast to the Tigris-Euphrates, which was open to invasions from many sides. It is therefore not surprising that the civilization that colonized the Nile was stable from before 3000 BCE until 300 BCE.

The desert climate is one of the reasons that we have many records from their time. Their favored writing medium was papyrus, and the dry climate was essential to its preservation.

1. Geometry

It is common to speak of geometry as originating with the need to reconstitute land markings washed away each year by the flooding of the Nile. We have seen that geometric knowledge considerably preceded the Egyptians. However, it reached an amazing level under them, and not just in surveying.

Their colossal architecture bespeaks extensive geometric knowledge. One feature of their construction—the sides of the pyramids, for example, or of some temples—is extremely accurate right angles. Those angles suggest that they knew the Pythagorean theorem. Knowing the theorem (more accurately, knowing its converse; see Exercise 2) tells you that if you have a rope 12 units long, and you stretch and form it into a triangle with sides 3, 4, and 5 units long, then the two short sides meet at right angles. If the unit is long, like 100 feet, then the right angle is precise.

> The suggestion is not hard evidence, given that there are other ways to build right angles. You can instead use this principle: In an isosceles triangle, the median from the apex (the segment from where the equal sides meet, to the midpoint of the third side) is perpendicular to the base. Thus, imagine a long rod with a mark at its midpoint, then ropes of equal length stretched from the ends of the rod to a common meeting point, then a third rope from the meeting point to the midpoint of the rod. Either way, though, we end up with evidence that is credible, even if not on paper, of understanding and experience in geometry.

Going further with surveying, the layout of tombs and temples suggests command of trigonometry. Another feature of the sides of the pyramids is extremely close alignment to the cardinal directions (north-south, east-west). Such alignment requires measurement of angles. Again we have only circumstantial evidence. If the Egyptians drew up trigonometric tables, they have not come down to us.

Numerous geometric formulas *have* come down to us. Some of them are simplifications indicating that the Egyptians were happy to take approximations instead of exact determinations. One formula has a version that says a circle of diameter 9 has the area of a square of side 8. In our notation, it says

$$8^2 = \pi\,(9/2)^2.$$

That amounts to taking the value of π as $256/81 \approx 3.16$. They took the same kind of liberty with other area and volume formulas.

One thing does appear to be certain: Whatever geometry they knew, the Egyptians did not do any geometric proofs. It seems that they accepted what experience, symmetry, or practical considerations indicated as true.

2. Numeration and Arithmetic

The Egyptians used a system of numeration based on decimal aggregates. Thus, there was a symbol for 1, a symbol for 10, one for 100, ... Then 257 would be represented by two hundreds, five tens, and seven ones. Such a system has the same property as Roman numerals: It makes addition easy and multiplication hard. (Think about Roman numerals and you will see that they include symbols for the decimal aggregates, together with intervening symbols for five times those aggregates.)

Still, with their civil engineering prowess, they must have possessed remarkable calculation skill. Their multiplication was dyadic: They used multiplication by 2, which is easy, and the property of the powers of 2 that any integer is the sum of distinct such powers.

> Consider the product 57 × 86. As a sum of powers of 2,
>
> 57 = 32 + 16 + 8 + 1.
>
> To produce that breakdown, observe that the biggest power of 2 that fits (does not exceed) 57 is 32. We have 57 – 32 = 25. Next, 16 fits 25, and 25 – 16 = 9. Next, 8 fits 9, and 9 – 8 = 1. Neither 4 nor 2 fits 1. That gives us the sum of powers.
>
> Now to multiply, we write 86 and double repeatedly:
>
> | | |
> |---|---|
> | 1 | 86 |
> | 2 | 172 |
> | 4 | 344 |
> | 8 | 688 |
> | 16 | 1376 |
> | 32 | 2752. |
>
> We then find the product by addition:
>
> 57 × 86 = 32 × 86 + 16 × 86 + 8 × 86 + 1 × 86
> = 2752 + 1376 + 688 + 86
> = 4902.

Another feature of their numeration and arithmetic was the exclusive use of **unit fractions**, fractions with numerator 1 (except for 2/3). Thus, for 2/5, they would have written 1/3 + 1/15. (Check that they match.) What advantage accrued to this habit is difficult to see. Such decomposition is not unique: 2/5 also equals 1/4 + 1/10 + 1/20. The only reasonably obvious algorithm to achieve the decomposition is the "biggest fit" process illustrated two paragraphs back.

> For 9/10, the biggest unit fraction that fits is 1/2; for 9/10 – 1/2 = 4/10, the biggest fit is 1/3; for 4/10 – 1/3 = 2/30, the biggest is 1/15; and we have
>
> 9/10 = 1/2 + 1/3 + 1/15.
>
> This method always works, because the numerators (in 9/10, then 4/10, then 2/30) necessarily decrease; see Exercise 6.

3. Algebra

The first evidence of what we would recognize as algebra is in the Rhind Papyrus, called also the Ahmes Papyrus. (**Boyer** identifies Ahmes as the author of the scroll around 1650 BCE, Rhind as a Scot who bought it in 1858.). It includes something like a workbook, in whose questions a quantity is to be determined based on information. One problem asks how to divide 100 loaves among five men so that the shares form an arithmetic progression and 1/7 the sum of largest three shares equals the sum of the smallest two. In our thinking, the problem comes down to simultaneous linear equations. A quadratic one, listed by both **Boyer** and **Struik**, asks for the sides of two squares, the smaller having side 10 less than 2/3 the side of the bigger, whose areas sum to 1000.

In the instruction that seems to have been the workbook's intention, there was no attempt to develop general methods applying to broad classes of problems—nothing like, say, the quadratic formula. Instead, the treatment was case by case: If the problem looks like this, then make this calculation. For the loaves problem, the "method of false position" was used. You simply take a guess at the result, then scale it up or down as required; see Exercise 5. The same method works, but rather slowly, in the problem with the squares.

4. Astronomy

One indication of the Egyptians' astronomical knowledge is their method for determining the coming of the flood. Recall our allusion to the Sun's annual trip around the sky. Its eastward motion makes the evening stars disappear into the dusk. It then reveals them to Sun's west, so that they become visible ahead of the dawn. The first day a star becomes visible in the pre-dawn sky is called the star's "heliacal rising." The heliacal rising of the most brilliant fixed star, Sirius (the "Dog Star"), was the harbinger of the Nile's annual flooding.

Notice that this Sirius connection implies the existence of a group with freedom to observe the sky, keep astronomical records, and draw conclusions from the information.

Such a group must have determined the length of the solar cycle, but it is interesting that they did not prescribe a calendar based on that length. Instead, the Egyptians used two calendars. One was strictly lunar and was used to time religious observances. The other was solar, used for civil purposes. The latter consisted of 12 months of 30 days each, plus five days that did not count as part of any month. (Why choose 12 months?) If it seems odd to have two calendars, visit a Jewish temple. Somewhere there, you will see an indication that the first day of the Hebrew month Tishrei, marking the start of the year, falls on some civil date; for example, it was September 9 in 2018.

A calendar covering 360 + 5 days is 1/4 day short of the year. If your calendar is short, then the seasons begin to move forward (later) within it. Imagine that your climate always makes the first snowfall happen on the first day of winter. If you call that date "Day 1" and your calendar is 1/4 day short, then four years later the first snow will fall Day 2; four years after that on Day 3; and so on. Over the course of a human life, the drift is barely noticeable—20 days if you manage 80 years. But over the culture's life of more than 2700 years, the seasons would have drifted 700 days, all the way through the calendar and almost again. It does not seem to have bothered the Egyptians.

Exercises II.A.4

1. a) Answer this question from **Boyer**: Which would have been more influential in the development of Egyptian geometry: surveying, or astronomy?
   b) Now consider: Would it have been the same for early humans?
   c) What would have been the first geometric figures to be studied systematically?
2. State the Pythagorean theorem. Then state its converse.
3. a) Find the product 394 × 53 by doing doublings (multiplications by only 2) and additions.
   b) Try Egyptian division: Calculate 95 ÷16 (ending up with integers and unit fractions) by doubling *and halving* the 16.
   c) What would happen if the divisor were 15 instead of 16? Why does that happen?
4. Use the biggest-fit algorithm to break 29/35 into unit fractions.

5. Try something like false position on the problem of the loaves. Begin by assigning the five men shares consisting of 1, 2, 3, 4, and 5 loaves. Those shares are in arithmetic progression, but do not satisfy the other requirements.
a) The sum of the lower two shares is 3, and 1/7 the sum of the upper three is 12/7. That is an error of 3 – 12/7 = 9/7. If you double the increment (the constant difference between consecutive terms), then the shares become 1, 3, 5, 7, 9, and the error drops to 4 – 3 = 7/7. Tripling the increment makes the shares 1, 4, 7, 10, 13 and the error 5 – 30/7 = 5/7. What multiple of the increment will eliminate the error?
b) Use the increment found from (a) to write down the five shares for which the sum of the lower two matches 1/7 the sum of the upper three. These five numbers do not sum to 100. Scale them up (multiply all by the same factor) to *make* them sum to 100.

6. a) Let m/n be a reduced non-unit fraction and $1/N$ the first unit fraction smaller than m/n. Prove that the fraction $m/n - 1/N$ has numerator smaller than m. In symbols, assume
$$1/N < m/n < 1/(N - 1)$$
(which precludes the possibility that m/n reduces to a unit fraction). Prove that
$$m/n - 1/N$$
has numerator smaller than m.
b) Argue why (a) proves that the biggest-fit method always produces a decomposition of a fraction into unit fractions.

Section II.B. The Babylonians

We noted that Mesopotamia (roughly modern Iraq) was always subject to invasion, so it is not surprising that numerous civilizations conquered the area and were in turn conquered. There have been important cities there for perhaps 8000 years. Babylon (roughly Baghdad) was one of them, and it was the capital for the civilization that dominated from around 1900 BCE to around 700 BCE. This group is our next interest.

Their writing medium was clay. They incised wedge-shaped (“cuneiform”) characters into clay tablets, which were then baked. The result was basically writing in stone, records that were extremely durable and have come down to us in wonderful quantity.

1. Geometry

With the Babylonians, we have the expected achievements in surveying and architecture. Their signature constructions were sawed-off pyramids called “ziggurats.” These were tall (and massive) buildings, most likely restricted to the priestly class.

They must have known the Pythagorean theorem, because they devised methods for producing integer-sided right triangles. A **Pythagorean triple** is a set of three natural numbers that make up the sides of a right triangle. In the language of algebra, a, b, and c (biggest) form a Pythagorean triple if
$$a^2 + b^2 = c^2.$$
One method is in Exercise 2a. Before we give the more complete one, let us discuss triples in general.

If a, b, and c all have a common divisor, then the triple is an overgrown version of a smaller one. Thus, 30-40-50 is just a 3-4-5 on steroids. So we confine our attention to **primitive triples**, those in which the three numbers do not share a common divisor. [We will allow ourselves the imprecision of saying “no common divisor” without adding “except 1.” The mathematical convention is to say “no *nontrivial* common divisor.” However, the locution “no nontrivial” is so troublesome to students that I will choose to be incorrect rather than clumsy.]

In a primitive triple, a and b cannot both be even. That would make c^2 even, which in turn would force c to be even (losing the primitivity); see Exercise 1a-b. But it turns out that they also cannot both be odd; see Exercise 1c. They have to have **opposite parity**; one even, one odd. Henceforth, when we write a primitive triple, we will assume that a is odd and b even.

For the second method from the Babylonians, we state a theorem that we will prove later.

Theorem 1. The numbers a, b, c form a primitive Pythagorean triple iff there are natural numbers u and v, having opposite parity and no common divisor, such that

$$a = u^2 - v^2, \qquad b = 2uv, \qquad c = u^2 + v^2.$$

Notice that the pair u, v satisfying those equations is *unique*. The triple determines u and v:

$$u^2 = (c + a)/2, \qquad v^2 = (c - a)/2.$$

Consider these examples:

u	v	$a = u^2 - v^2$	$b = 2uv$	$c = u^2 + v^2$
2	1	3	4	5
3	1	8	6	10
3	2	5	12	13
4	1	15	8	17
4	3	7	24	25

The four shown in black are perhaps the most familiar to students. The one in red shows why u and v need opposite parity; if they have like parity, then all three numbers in the triple are even.

Like Egyptian geometry, Babylonian geometry had no hint of proofs. There is no surviving evidence that they thought the Pythagorean theorem needed geometric proof or that the two methods for producing triples needed numerical/algebraic justification.

Exercises II.B.1

1. a) Show that the square of any even number is a multiple of 4.
   b) Show that the square of any odd number is 1 more than a multiple of four.
   c) Use (a) and (b) to show that two odd squares cannot add up to a square.
2. Show that:
   a) If $m > 1$ is odd, then
   $m, \quad (m^2 - 1)/2, \quad$ and $(m^2 + 1)/2$
   form a Pythagorean triple.
   b) The triple in (a) is primitive.
   c) In this Babylonian way of constructing triples, m gives the shortest side.
   d) Not all primitive triples come from this method.
3. Find two Pythagorean triples in which one of the sides is 37. (They will necessarily be primitive, because 37 is prime. Theorem 1 then says those are the only two, one from 37 as a difference of squares, the other as sum of squares. Does Exercise 2 provide a third one?)

2. Arithmetic

Babylonian numeration was a remarkable advance: It was the first to use a **place-value system**.

Recall that our numeration uses the principle that every integer has a unique representation as "a polynomial in 10" with the **digits** 0-9 as coefficients. Thus,

$$2011 = 2 \times 10^3 + 0 \times 10^2 + 1 \times 10^1 + 1 \times 10^0.$$

In this scheme, the **base** 10 is not essential. We could as well have used 5 or 20 or 2, any natural number but 1, with the digits corresponding to that base.

The Babylonians chose **sexagesimal**, base 60, numeration. Thus,

(decimal) 20110 = $5 \times 60^2 + 35 \times 60^1 + 10 \times 60^0$.

They would have represented that expression by marks and aggregates for (our) 5, 35, and 10. For our purposes, let us agree to use

#5 #35 #10

to signify their expression. You can see the need to abbreviate the coefficients with aggregates; the alternative is to invent 60 symbols for the “digits” 0 to 59. On the other hand, a place-value system always makes both addition and multiplication receptive to the kinds of algorithms we learn for decimal numeration. Separately, the many divisors of 60 create shortcuts for some multiplications.

> Recall that the factoring 10 = 5 × 2 implies that you can replace multiplication by 5 with division by 2, which is easier, followed by multiplication by 10, which requires only moving the decimal point. Thus, you can do 187 × 5 by taking half of 187, or 93.5, then deciding where the decimal point goes. In the same way, 5 × 12 = 60 implies that
>
> (#5 #35 #10) × (#12) = #5/5 #35/5 #10/5 #0
> = #1 #7 #2 #0.
>
> (Check that
> 20110 × 12 = $1 \times 60^3 + 7 \times 60^2 + 2 \times 60^1 + 0 \times 60^0$.)

The numeration departed from strictly positional when it came to terminal zeroes. When a place was missing, as in decimal 2011, the Babylonians would use a mark or an empty space to indicate the missing power of 60. (Eventually, they did adopt a symbol for zero.) However, that last expression

#1 #7 #2 #0 might instead be written #1 #7 #2.

That would leave it to the reader to determine from the context whether

$1 \times 60^3 + 7 \times 60^2 + 2 \times 60^1$ or $1 \times 60^2 + 7 \times 60^1 + 2 \times 60^0$

(or even $1 \times 60^1 + 7 \times 60^0 + 2 \times 60^{-1}$) were meant.

3. Algebra

The records of Babylonian algebra are like the Egyptian. There are what seem to be books intended to instruct students in the solution of problems. The problems tended to have such practical orientation as the division of collections of objects or bodies of land. Their solutions are case-by-case: Faced with this situation, you do this and that. Still, their *ad hoc* approaches must have pointed to patterns, and they managed with such approaches to solve (approximately) some kinds of quadratic, even cubic, equations.

The Babylonians shared the Egyptian willingness to accept approximate answers. They did not, for example, have representations for square roots. A reference to √40 would have referred to a rational approximation. According to **Boyer**, they created this **square-root algorithm**:

Given a positive x, not necessarily integral, let y be any positive guess at $\sqrt{x}$; then

$$\sqrt{x} \approx (y + x/y)/2$$

is a refined guess.

Let us approximate √40, which we know is around 6.

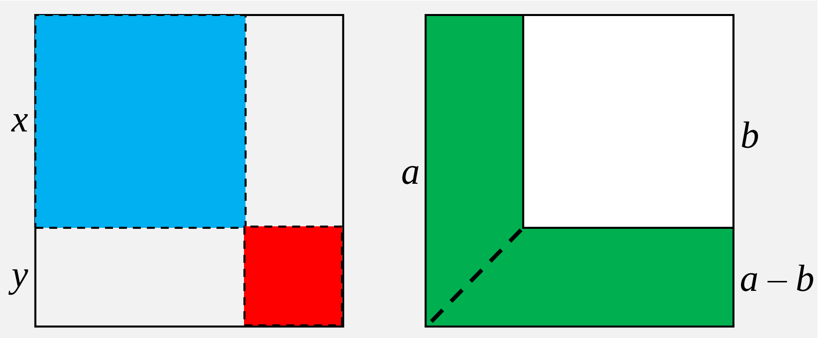
> Begin with the guess 10 (because it divides 40). The algorithm names only two steps:
>
> Take 40/10 = 4, then average 10 and 4 to get 7.
>
> Notice that 10 and 4 are on opposite sides of the known value, their average is too high, and the average is closer than either of the two; see Exercise 5. More important, the algorithm is **recursive**; the result can be fed back to it to produce further refinement. Thus,
>
> 7 has the partner 40/7 (which is less than 6), and their average is 89/14.
>
> The new estimate is again greater than and closer to √40. ($7^2 = 49$, $[40/7]^2 \approx 33$, but $[89/14]^2 \approx 40.4$.)

One feature of their algebra is a predilection for “geometric algebra,” pictorial renderings of relations we would consider algebraic.

To us,

$$(x + y)^2 = x^2 + 2xy + y^2$$

is an algebraic identity. The Babylonians would, first off, not have thought of a “square” as the result of multiplying a number by itself. They would have imagined the quadrilateral. In the left half of the figure at right, we have a square of side $x + y$. The dashed lines clearly break it up into a square (blue) of side x, another (red) of side y, and two rectangles of sides x and y. The figure portrays the identity.



In the same way, the right half of the figure illustrates

$$a^2 – b^2 = (a + b)(a – b).$$

There, the L-shaped green area is $a^2 – b^2$, and it can be cut along the heavy dashed line into two pieces that fit back together to make a single rectangle of sides $a – b$ and $a + b$. See also Exercise 6.

Exercises II.B.3

1. What is the etymological meaning of “Mesopotamia”? of “mathematics”?
2. (Boyer) What does “geometry” mean etymologically? Is the use of the word justifiable, given the subject’s development?
3. Name some Mesopotamian contributions to our mathematics.
4. The areas of two squares sum to 895. The side of one square is 10 less than two-thirds the side of the other. Use the method of false position to approximate the sides of the two squares to within 0.01. You may use the arithmetic functions of a calculator. Keep track of how many calculations are needed, and how you can cut down on them.
5. a) Apply the square-root algorithm to $\sqrt{3}$, beginning with the estimate $\sqrt{3} \approx 1$ and going as far as three “new estimates.”
   b) Check that each new estimate and its partner straddle the exact root. Explain why that happens.
   c) Check that each new estimate is rational. Why is that always so?
   d) Check that each new estimate is too big. Show that this must happen.
   e) Show algebraically that each new estimate cuts the error by more than half (is less than half as far off as the previous estimate was. You can show it via calculus as well.)
6. Draw a figure to illustrate the identity
   $$(x – y)^2 = x^2 – 2xy + y^2.$$

4. Astronomy

The Babylonian record in astronomy is remarkable, particularly in their tracking of the Sun. They used their observations to derive antiquity’s best estimate of the year.

We have referred to the Sun’s apparent trip around the sky over the year. During that trip, the Sun traces out a path among the stars, a great circle called the “ecliptic.” No doubt viewing it as the path of life, the Babylonians organized the star groups along it into [poor] representations of living creatures.

The resulting constellations include such real beings as a pair of twins and a scorpion, and such mythical ones as a “sea goat.” [That menagerie of creatures was later named by the Greeks. Doubtless you know the name: “Zodiac,” from the Greek word for “life.” Think of “zoology.”]

Because Earth’s rotation axis points about 66.6° off the plane of the ecliptic, the “Celestial Equator” (the great-circle projection of Earth’s equator into the sky) is at the complementary 23.4° angle to the ecliptic. The Celestial Equator and ecliptic meet at two points, called the “equinoxes” (from the Latin for “equal nights.” On the day the sun reaches an equinox, the daylight and night are [theoretically] equally long.) Around March 21, the Sun reaches the “vernal” (spring) equinox, in The Fishes. It is then headed northeast in the sky. Around June 21, the Sun reaches its northernmost point on the ecliptic, the “summer solstice,” near The Twins. That position gives the Sun high elevation in the daytime sky and longer visibility (the long days). The combination of increased elevation and time above the horizon accounts for our summer. Thereafter the Sun’s path starts east, then curves to southeast, to cross the Celestial Equator around September 21. That crossing is the “autumn equinox,” near The Virgin. After that, the Sun continues southeast. It arrives around December 21 at its most southerly point, the “winter solstice,” between The Scorpion and The Archer. That southerly location gives us short days with the Sun low in the sky, and therefore winter.

[For all those dates and sky locations, you have to say “in our era.” They vary through a cycle of about 26,000 years, in a way and for a reason that we will discuss later.]

The Babylonians measured the year by counting the days between spring equinoxes hundreds of years apart. By careful observation through viewing holes or tubes—or by measuring shadows—you can find the directions of extremes: the angles of elevation of summer’s highest sun (at the summer solstice) and winter’s lowest midday sun (winter solstice); or the directions of summer’s northernmost and winter’s southernmost sunsets; or the same for sunrises. (Many cultures built observatories—temples, really—oriented to, or holding markers indicating, the directions of those “solstitial” rises and sets. Stonehenge in England and the Sun Temple at Machu Picchu in Peru are two famous examples.) Halfway between the directions in any of those pairs lie the two equinoxes. If you track the equinoctial days over a span of 600 of them, even with an error of two days on each end, you measure the year to within 4/600 day (24/3600 day in sexagesimal, less than 10 minutes).

> Recall our observation that one human can record a hundred Full Moons. Obviously the same is impossible with hundreds of equinoxes. On the other hand, you can see how a society spanning more than a millennium managed to document them.
>
> For the Babylonians, the estimate of the span between consecutive spring equinoxes was
> #6 #5 **PT** #14 #33
> days. Here we have invented **PT** to signify the “sexagesimal point.” Thus, the numeral above represents
> $6 \times 60^{1} + 5 \times 60^{0} + 14 \times 60^{-1} + 33 \times 60^{-2} =$ decimal $365 + 873/3600$.
> Notice that the number is less than
> $365\frac{1}{4} = 365 + 900/3600$.

Chapter III. The Greeks

The Greeks were a kind of counterpoint to the Babylonians. That statement should not surprise, because the Greeks and Babylonians were separated in time, space, and culture. The Babylonian time ended by 700 BCE, before the Greek began. Mesopotamia was largely landlocked east of Turkey and owed its existence to two rivers; Greece was a peninsula, daughter of the sea. Babylon was ruled by kings; Greece came to be dominated by a merchant class that invented democracy. For us students of math and science, the Greeks are the natural next interest.

Section III.A. Geometry

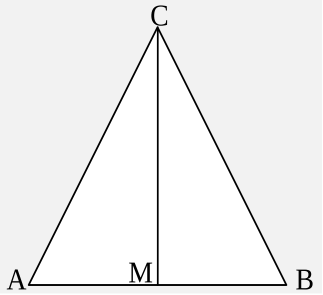
1. Thales

The study of Greek math begins with Thales. He was a merchant from Miletus. (Miletus is on the western margin of Turkey. We could reasonably say “Aegeans” in place of “Greeks.” Greek civilization included modern Greece on the western coast of the Aegean, parts of Turkey on the eastern, and the islands in between.) **Boyer** mentions evidence that Thales lived 40 years either side of 600 BCE.

With Thales, we have the first instance of mathematical knowledge ascribed to a specific person. The attribution is apocryphal; **Boyer** has a wonderful paragraph (“There is no document ...”) that illustrates how slippery history can be. What is sure is that Thales marks the beginning of “demonstrative geometry.” Nothing before Thales indicates that people thought it necessary or useful to “prove theorems.” Nobody before put forth arguments purporting to demonstrate that the truth of one statement follows from the truth of others (whose truth was perhaps demonstrated or accepted before).

To see what we mean, look at an example. Among the theorems Thales is said to have proved are the equality of base angles of an isosceles triangle, equality of vertical angles, measure of an angle inscribed in a semicircle, and congruence of triangles based on AAS. Look at the following argument for the base-angles statement.

> In the figure at right, we have triangle ABC with sides AC and BC congruent. [I was trained to say the sides are “equal.” I will try to adhere to the usage from *after* the Civil War; please forgive any occasional lapse.] We draw the median from C to the midpoint M of AB, creating the new triangles AMC and BMC. In those two, AC ≅ BC by assumption, AM ≅ BM by construction, and CM is identical to itself. By SSS, triangles AMC and BMC are congruent. Therefore their corresponding parts match, making angle A congruent to angle B.



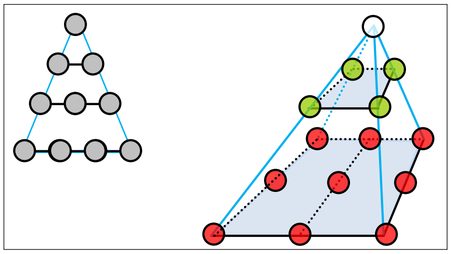
Notice that the argument says that the base-angles statement is true *because* the SSS principle is true. There is no record that SSS had been proved before, but it may have been accepted as supported by long experience. The physical principle that the triangle is rigid is a manifestation of a geometric one, that if the sides of a triangle match up with those of a second, then the triangles have the same size and shape.

2. The Pythagoreans

Not much is known about Pythagoras, except that he died some fifty years after Thales (under whom he may have studied), around 500 BCE. By his time, though, it was possible to make a living by establishing a “school.” If you acquired a reputation for knowledge or wisdom, you could gather followers around you who would pay or otherwise support you, in exchange for getting to learn from your thinking. It is the disciples, the Pythagoreans, whom we focus on.

The Pythagoreans formed a secret, mystical society. They were given to all sorts of beliefs about the supernatural power of numbers. This focus on number will make us come back to them later. [Plato's school was an actual, physical academy, whose gate was said to bear a sign to the effect of, "Let no one ignorant of mathematics enter here." That sounds like a salute to the glory of mathematics. My colleague Ethan Akin points out that the Greek name for mathematics amounted to "geometry." Accordingly, he says, the sign was actually a suggestion that the number-mad Pythagoreans might take their talents elsewhere.] Still, they were Greeks, and therefore well versed in geometry.

a) figurate numbers



Typical of the Pythagoreans was the idea of **figurate numbers**. In the left half of the figure at right, we have 1 gray ball, then 1 + 2, then 1 + 2 + 3, then 1 + 2 + 3 + 4, formed into triangles. Accordingly, the numbers 1, 3, 6, 10 ... are called **triangular** numbers. The Pythagoreans produced the equivalent of the formula

$$1 + 2 + ... + n \;=\; n(n + 1)/2.$$

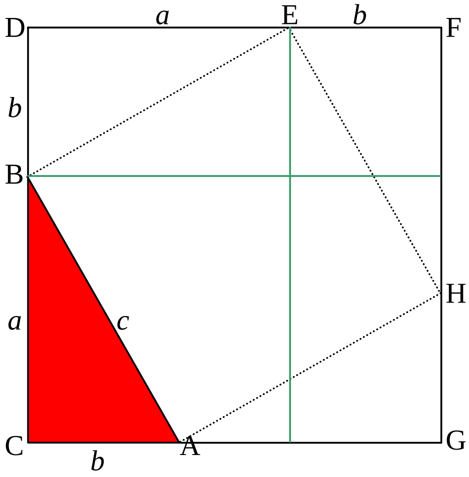
In the right side of the figure, which tries to show three dimensions, we form 1 (white) ball, then 1 + 4 (4 greens), then 1 + 4 + 9 (9 reds) into tetrahedra. We may call the resulting 1, 5, 14, ... **pyramidal** numbers. Naturally, the Pythagoreans knew

$$1 + 4 + 9 + ... + n^2 \;=\; n(n + 1)(2n + 1)/6.$$

[We typically meet these formulas in the calculus introduction to integrals. They are worth memorizing.]

b) proving the theorem

From the Pythagoreans, we have the first *proof* of the theorem that bears their name. We will give a proof here that relies entirely on a picture. (Compare Exercise 2.) The proof is not remarkable; over the centuries, people would produce original proofs for the purpose of showing off their erudition.



At right, we start with right triangle ABC (filled in red). It has AC of length b horizontal, BC = a vertical. We extend CB upward a length b to point D. At D, we draw the horizontal rightward, a length a to point E, then a further length b to point F. At F we draw the vertical downward, meeting the extension of CA at G. By construction, quadrilateral CGFD has four right angles. It has congruent adjacent sides CD and DF, both $a + b$ long. Therefore it is a square. That means AG has to be a, and if we pick H to be distance a down from F, then necessarily HG = b.

The green horizontal at B and vertical at E break up CGFD into the square on BC, the square on EF (the same size as that on AC), and two rectangles. (If we were talking algebra, we would say this merely reflects $[a + b]^2 = a^2 + 2ab + b^2$.) Therefore the area of CGFD is the sum of the square on BC, the square on AC, and two rectangles.

The dotted lines BE, EH, and HA create three corner triangles that are congruent to the original ABC (right triangles, legs a and b). That tells us that quadrilateral AHEB has four congruent sides, c long. It also tells us that angle HAG matches angle ABC. Since angles ABC and CAB add up to 90°, it follows that angles HAG and CAB add up to 90°. That leaves the same 90° for angle BAH. Hence AHEB is a rhombus with a right angle; it is the square on AB. It follows that the area of CGFD is the sum of the square on AB and four triangles.

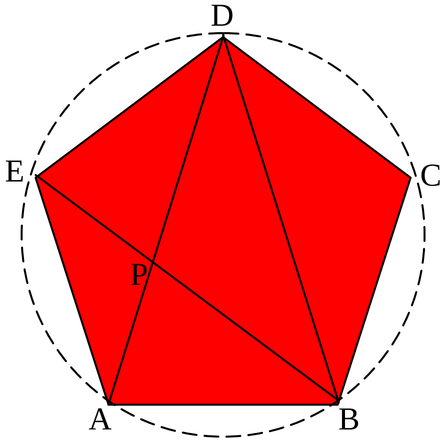
> We now know that
> area CGFD = square on BC + square on AC + two rectangles
> = square on AB + four triangles.
> Clearly the two rectangles are congruent, and the rectangle at upper left is twice the size of the triangles. Therefore the two rectangles sum to the area of the four triangles. We conclude that the square on the hypotenuse is the sum of the squares on the other two sides.

c) pentagons and the section

Our final sample of Pythagorean geometry involves regular pentagons. Below we draw regular pentagon ABCDE, filled in red. We add the circumscribed circle, dashed. Then we add the **diagonals** AD, BD, and BE, which meets AD at P. The Pythagoreans discovered that

AD/PD = PD/AP.

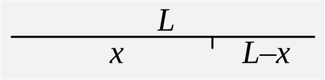
In words, any two diagonals so intersect that for each, the whole is to the longer piece as the longer piece is to the shorter. (Prove it in Exercise 4a, based on the information below.)



> It is a geometric fact that every regular polygon can be inscribed in a circle. If you draw the perpendicular bisectors of adjacent sides, like AB and BC, the bisectors intersect at a point equidistant from A, B, and C. You can accept from symmetry, or prove with some effort, that the intersection is actually equidistant from *all* the polygon's vertices. Accordingly, the circle centered there and reaching one vertex reaches them all.
>
> The circle is important because it gives us angle measures. First, each angle of the pentagon is inscribed in an arc 3/5 of the circle. That makes each angle measure
>
> ½ × 3/5 × 360° = 108°.
>
> (An even more familiar principle says that the exterior angles of a polygon share 360°. It implies that the pentagon's exterior angles are 72° apiece, leaving 108° for the interior angles.) Second, the angles of triangle ABD are all inscribed. They have to measure 36-72-72. In triangle DEP, the angles at D and E are inscribed. They have to be 36° and 72°, respectively, leaving 72° for the angle at P. All the tall, narrow triangles, like BPA, have angles 36-72-72. By analogous reckoning, the short, wide triangles, like APE, have 108-36-36 angles.
>
> [Leaving out the other diagonals makes the proof in Exercise 4a easier. Add the remaining two diagonals. You will see the (**sign of the**) **pentagram**. Even if you never saw the original *Wolf Man* movie (Lon Chaney Jr., from Universal Studios), you can imagine the mystic properties the boys would have ascribed to this figure. Separately, you will see a smaller hexagon inside the original; go to Exercise 4b.]

The relationship in which the whole bears to the bigger piece the same ratio that the bigger bears to the smaller is called the **golden section (**or **golden mean** or **golden ratio**.) The ratio has many properties, including one whose basis appears to be aesthetic. [In other words, I do not know a mathematical basis for it.] That property is the idea that the most pleasing proportion for a rectangle—what we now call the "aspect ratio"—prevails when the width bears the golden ratio to the height. That is the same as saying that width + height bears the golden ratio to the width.

We are going to do two things with the ratio: First we evaluate it, then we construct it.



Let us **section** a line segment of length L into pieces x and $L - x$, so that

$L/x = x/(L - x)$.

First, divide numerators and denominators by L, to make the proportion

$1/(x/L) = (x/L)/(1 - x/L)$.

If we substitute y for x/L, we see that the original length L is irrelevant; the proportion amounts to

$1/y = y/(1 - y)$.

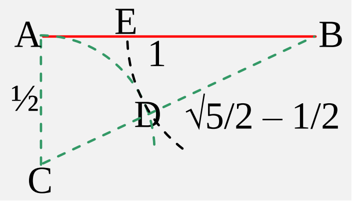
The resulting quadratic is yours to solve. Its positive solution is

$y = (\sqrt{5} - 1)/2 \approx .62$.

This is informative; it says that in the section, the longer piece is about 0.6 of the whole. Remember, though, that it was $L/x = 1/y$ we named "the ratio":

$1/y = 2/(\sqrt{5} - 1)$
$= (\sqrt{5} + 1)/2$ (Prove that equality.)
≈ 1.62 (Is the ".62" a coincidence?)

For the construction of the ratio, we show one way to "section" a segment.



At right, start with segment AB (red), whose length we call 1. Along the perpendicular at A, lay off half the length to C. Then BC has length

$\sqrt{(1^2 + [½]^2)} = \sqrt{1¼} = \sqrt{5}/2$.

Draw the arc centered at C through A, meeting BC at D. That leaves

BD = $\sqrt{5}/2 - 1/2$.

The black dotted arc, centered at B through D, sections AB at E.

[Throughout the book, after a construction has been described in the text or an exercise, you may *invoke* it. Thus for example, you may say "Section this segment" without going through the description above of how to do it.]

Exercises III.A.2

1. (Boyer) Prove one of the theorems ascribed to Thales. Would he have reasoned the way you do?
2. Prove *by means of a picture* that:
   a) $1 + 2 + ... + n = n(n + 1)/2$.
   b) The sum of consecutive triangular numbers is a square (like 6 + 10 = 16).
3. How many diagonals does a regular pentagon have? How many in an *n*-gon?
4. a) Prove that the diagonals of a regular pentagon "section" each other. (Hint: The question asks you to prove a proportion. In geometry, there is just one concept related to proportionality.)
   b) The diagonals bound a little top-down pentagon, clearly regular, within the original one. Show that the sides of the original bear the square of the golden ratio to those of the inner. What happens if you draw the diagonals of the inner one?
5. Describe the straightedge-and-compass construction of a regular pentagon:
   a) given one side. (Hint [from one of my students]: "Unsection" it. To start the sectioning above, you build a right triangle whose hypotenuse has the length of the pentagon's diagonals. Build the diagonals to the given side and work from there.)
   b) or instead, given one diagonal. (Hint: Section it.)
   c) or instead, given the circumscribed circle (hereafter **circumcircle**). (Hint: Start to section the radius; establish that the right triangle has an 18° angle at the circle's center.)

3. Constructions

The century from 500 to 400 BCE brought tremendous achievement in Greek geometry, but we will focus largely on constructions.

a) constructions

A “construction” is a drawing process that produces a model of a geometric figure. At the most elementary level, we can use a compass to construct parts of circles, and a straightedge—a ruler with no markings on it—to construct parts of lines. Exercise 1 lists some of the earliest constructions students encounter in the elementary geometry course. Exercise 2 has some other, harder elementary ones, and Exercise 3 gives comparatively advanced ones.

The rule that you must do constructions armed with just straightedge and compass comes from later, but it was already in Euclid’s writing around 300 BCE. Even now people do not always agree on it, and you can make challenges under other rules; see Exercise 4. For that matter, there is not universal agreement on what makes a compass. To some, a compass is the familiar kind that holds its setting by friction or by a screw. To others, it is one of those classroom tools with a loose hinge, so that its arms flop together or apart if you hold just one of them. We will stick with the standard rule and instruments.

Notice that we distinguish between “elementariness” and “difficulty.” A question is **elementary** if you can understand it. Otherwise, it is **advanced**. A question is **difficult** if you are unable to answer it; otherwise, it is **easy**. The question of constructing a regular pentagon, Exercise 5 in the previous section, is an example of an elementary but hard construction. Both ideas are evidently relative: What is elementary to a graduate student is advanced to a ninth-grader, and often a hard question comes to seem easy after you have seen it or a close relative answered.

b) the three ancient problems

In that fifth century, Anaxagoras popularized three wonderfully elementary constructions. One was the problem of **squaring the circle**.

Problem 1. Given a circle, construct a square of equal area.

Instead of addressing the problem, put it into perspective with the problem of squaring a *rectangle*.

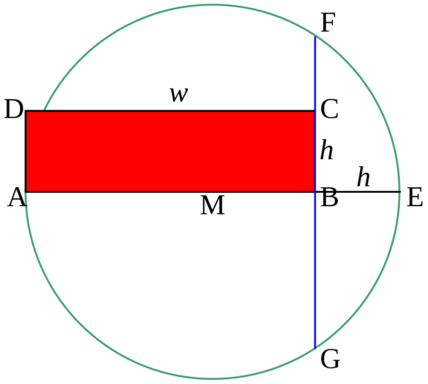
In the figure at right, we have rectangle ABCD in red, w wide by h tall. We extend the width AB by h to point E, then find the midpoint M of AE. We draw in green the circle centered at M of radius AM = ME. Then we build the blue perpendicular to AE at B. It meets the circle at F and G. We employ two properties of chords in circles. First, if a chord like FG is perpendicular to the diameter AE of a circle, then the diameter bisects the chord. That implies FB = BG. Second, if any two chords meet inside a circle, then the product of the (lengths of the) pieces of one equals the product of the pieces of the other. That makes

$$AB \times BE = FB \times BG,$$

or

$$wh = (FB)^2.$$

In other words, FB is the side of a square of area wh. We then build the square (not shown) as per Exercise 1d, and we have squared the rectangle.



The other famous problems were **trisecting an angle** and **doubling a cube**.

Problem 2. Given an angle, construct the rays that partition it into three congruent angles.

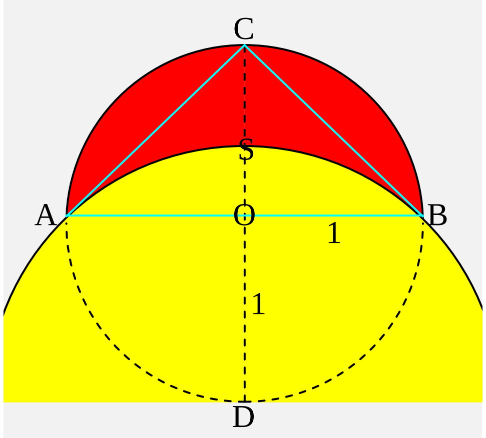
Problem 3. Given a cube, construct (one edge of) a cube with twice the volume.

Contrast Problem 2 with Exercise 3a, which implies that breaking a *segment* into three or more congruent pieces is advanced but doable. Similarly, contrast Problem 3 with Exercises 2a and 5b, which together imply that you can duplicate or triplicate or more any square.

The contemporaries of Anaxagoras knew how to square any straight-sided figure (Exercise 5d). It was natural that they should turn to the most elementary curved figure. But squaring the circle turned out to be hard. Really hard; nobody could do it. Hundreds of years later, still nobody had broken the problem. It was enough to make geometers wonder whether the construction was impossible.

Now that brings up a truly advanced question. What does it mean to say a task is impossible? Obviously you can prove that a construction is possible by doing it. How could you possibly prove impossibility? We will come back to the question. Meanwhile, it is worth reflecting on a numerical impossibility we already met: According to Exercise II.B.1:1c, you cannot find two odd numbers whose squares add up to 1,000,000 = 1000^2. See also Exercise 6 here.

The difficulty was especially intriguing because Hippocrates (not the medicine man, but like the physician, a native of the islands) managed to square one particular region with curved sides. A **lune** is the region inside one circle and outside an overlapping, bigger circle. In the figure at right, we draw the circle centered at O, filled in red, with a radius that we set at 1 unit. AB and CD are perpendicular diameters of it. In front of that circle, we draw the second circle (yellow), centered at D and reaching A and B. The lune is the visible red part. Hippocrates squared that lune by showing that it has the same area as triangle ABC.



Draw in blue the chords AC and BC in the first circle and AB, which we now view as a chord, in the second. The region bounded by a chord and the shorter arc of its circle is called a **segment** of the circle. It is clear that AC and BC span a quarter of the first circle. Therefore they bound congruent segments. Because angle ADB is a right angle (Why?), AB spans a quarter of the second circle. Consequently, the segment it bounds in the second circle is **similar** to those of AC and BC in the first.

Here we are extending the concept of similarity to regions other than triangles, but the properties that go with the concept match what happens in triangles. When you have similar segments of circles, their corresponding linear elements are proportional, and those (line segment) elements are in the same ratio as the two radii. For example, you can see that the radius DB of the second circle is √2; the ratio of the radii is √2/1. For the chords, AB has length 2, AC has length √2 (Why?), for a ratio of 2/√2 = √2. In the same way, the **sagitta** of AB—the perpendicular distance OS from the midpoint of the chord to the arc—is obviously √2 – 1. It is less obviously 1 – √2/2 for AC or BC. (Check that last measurement by drawing the radius that bisects angle AOC. Then check the ratio between √2 – 1 and 1 – √2/2.)

More important for similar circular segments, the ratio between their areas is the *square* of the ratio between the linear parts. It follows that the area of the segment bounded by AB is twice the area of the segment bounded by AC or BC. In other words, the AB segment has the area of the other two combined. (It was Hippocrates who first proved the statement about the ratio of the areas. The statement's restriction to semicircles is itself important; it implies that the (ratios of) areas of circles are as the squares of the radii.)

> Now see why the area of the lune is the area of right triangle ABC. The lune consists of the two smaller segments and the part of triangle ABC above the larger segment. Therefore
>
> area of lune = areas of the smaller segments + area of the triangle – area of the larger segment.
>
> (You should consider that this relation depends on the fact that the larger circle does not stick out above the triangle. The reason is that AC and BC are tangent to the larger circle; explain why.) Since the smaller segments add up to the larger, the lune is the size of the triangle. If we build the square whose sides are OC and OB, then we complete the **quadrature**, the squaring, of this lune.

Exercises III.A.3. This is a titanic exercise set. It will profit you to *read* all the exercises. However, discretion being the better part of valor, it will be wise to do only a selection of them.

1. Describe the construction of:
   a) a copy of a line segment (a segment equal in length to a given one, or to the distance between two given points);
   b) the perpendicular bisector of a given segment;
   c) the perpendicular at a specified point of a given segment;
   d) a square having a given segment as one side;
   e) the perpendicular to a given line from a given point off the line;
   f) the bisector of a given angle;
   g) a copy of a given angle, the copy having a given segment or ray as one side.
2. Describe the construction of:
   a) a square twice as big as a given one (It is easy to make one twice as wide and twice as tall. But, a square being two-dimensional, the natural measure of size is *area*. This question wants a square with twice the area.);
   b) a square half as big as a given square;
   c) the circumcircle of a given triangle;
   d) the circumcircle of a given regular polygon.
3. Describe the construction of:
   a) the points that divide a given segment into a specified number of congruent pieces;
   b) the two tangents to a given circle from a given point outside the circle.
4. With only a compass, construct two circles that intersect at right angles (circles with perpendicular tangents at their two points of intersection).
5. Describe how to:
   a) square a triangle. (Hint: First enclose the triangle in a rectangle.)
   b) construct a square whose area is the sum of two given squares.
   c) construct a square equal to the difference of two given squares.
   d) square a polygon.
6. a) Prove that if a quadrilateral is inscribed in (has all four vertices on) a circle, then its opposite (nonadjacent) angles are supplementary.
   b) Prove that it is impossible to inscribe a parallelogram in a circle, except for a rectangle.

4. Eudoxus

The years around 400 BCE were eventful for Athens. The city fell to Sparta in 404, Socrates died in 399, and Eudoxus (aged maybe 21) moved there around 387. The works of Eudoxus are about as extensive and brilliant as those of Archimedes, who attributed many results (to which he brought original insights) to Eudoxus. We will get into discoveries about numbers later. Here we look at Eudoxus's most famous study in geometry, the method of exhaustion.

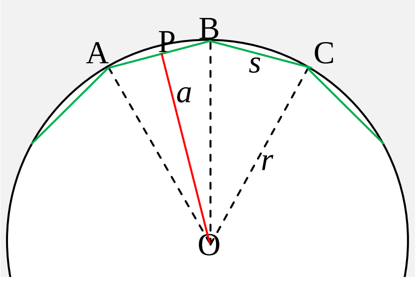
a) the question

Long before Eudoxus, the Greeks knew that you can approximate the area of the circle by resort to inscribed regular polygons, for which they could compute areas. The idea led to a formula.

In the figure at right, we have a circle of radius r centered at O. We label three consecutive vertices A, B, C of the inscribed regular n-gon (green outline). Each of its sides has length s. The altitude OP (red) of triangle AOB is called the **apothegm** of the polygon, as is its length a. The area of triangle OAB is $as/2$. The total area of the polygon is therefore

$$n\,as/2 \;=\; a(ns)/2 \;=\; aP/2,$$

with P being the perimeter of the polygon.



Picture now a regular polygon of a million sides. You see that it so hugs the circumcircle that its perimeter is practically the circumference C, its apothegm practically the radius r, and its area practically the area A of the circle. From

$$A \;\approx\; \text{area of polygon} \;=\; aP/2 \;\approx\; rC/2,$$

we conclude that the area of the circle is given by

$$A = rC/2.$$

To reconcile that formula with the familiar one, think in terms of Hippocrates's discovery about corresponding parts. Given the million-gon we already have, suppose a second one is inscribed in a circle of radius r^* and circumference C^*. The constituent triangles of the two million-gons are similar, because each triangle is isosceles and has an apex angle of $360°/10^6$. Therefore the second apothegm a^*, side s^*, and perimeter $P^* = 10^6 s^*$ are in proportion to the radii:

$$a^*/a \quad = \quad s^*/s \quad = \quad P^*/P \quad = \quad r^*/r.$$

To the extent that C^* and C are indistinguishable from P^* and P, the last equality gives

$$C^*/C \;=\; r^*/r.$$

Rewrite that as

$$C^*/r^* \;=\; C/r.$$

It says that in any two circles, the ratio between circumference and radius is the same. Write that constant as 2π, even though the "π" symbol came into use 2200 years after Eudoxus. Then we have

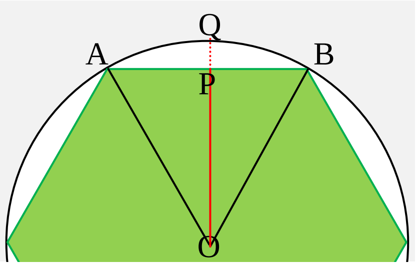
$$C = 2\pi r \qquad \text{and} \qquad A \;=\; rC/2 \;=\; \pi r^2.$$

(Separately, without resort to π: The areas of the similar million-gons are in proportion to the squares of the radii. It is reasonable to figure that this reasoning is how Hippocrates concluded that the areas of the *circles* are in the same proportion.)

b) the method of exhaustion

What the method exhausts is the area of a circle. It is intuitively clear that you improve the approximation by applying polygons of more and more sides. The key to the **method of exhaustion**—what made it conclusive—is that if you begin with one inscribed regular polygon and *double* the number of sides repeatedly, then the resulting polygons *exhaust* the area. That is, the area of the circle unoccupied by the polygons—what we will begin to call the **error**–gets closer and closer to zero.

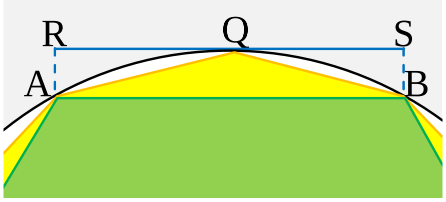
To prove that the approximation approaches the exact area, we start with a hexagon. The hexagon has the desirable property of being easy to construct.



In the figure at right, we visualize the constituent triangle AOB of the inscribed regular hexagon (green). The angle AOB is 360°/6. Therefore the triangle is equilateral, and the side AB of the hexagon equals the radius.

For that reason, the hexagon is the easiest polygon to inscribe. To do so, we draw the circle and keep the compass setting. Put the compass point at A and mark off B with the other leg. Then put the compass point at B, mark off the next vertex, and continue around back to A. You thereby determine the six vertices of the hexagon.

Now add to the figure the apothegm OP (red), extending it to Q on the circle. Because OP is the altitude in an isosceles triangle, it bisects angle AOB. Necessarily, Q bisects arc AB. That means A, Q, and B are consecutive vertices of the inscribed regular **dodecagon** (12-sider).



Next, at right we magnify the top part of the previous figure. We show the dodecagon (yellow) peeking out from behind the hexagon. It is clear that the dodecagon takes in some area of the circle that the hexagon missed; it gives a better approximation to the circle. To see how much better, draw at Q the tangent (blue), which is parallel to AB. (Why?) Add also the perpendiculars (dashed) from A and B to the tangent, meeting the tangent at R and S. From the top of the circle, the hexagon leaves out the (circle's) segment bounded by AB. From that error, the dodecagon eats up the area of triangle ABQ. Triangle ABQ, like any triangle inscribed in a rectangle, has half the area of rectangle ABSR. The area of ABSR exceeds the area of the AB segment. Therefore the dodecagon recovers more than half of the area the hexagon leaves out. The dodecagon's error is less than half the hexagon's error. (Do Exercise 1 to illustrate the calculations that follow from this idea.)

The above argument applies to any number of sides and its double. Each doubling of the number of sides reduces the error by more than half. We infer that the areas of the (3×2^n)-gons approach the area of the circle.

c) the need

This more-than-half worry is a response to the demands of Zeno's logic. Zeno was a philosopher living about 50 years before Eudoxus. He argued that mathematical thinkers were resorting to unreliable logic. He was especially wary of (what we call) infinite processes (like exhaustion) that indefinitely reduce quantities (like the method's error). To make his point, he propounded "Zeno's paradoxes." A **paradox** is just a contradiction. Paradoxes are typically presented as seemingly sound arguments that lead to a conclusion contrary to experience or to another apparently sound line of reasoning. The most famous of his paradoxes, which has multiple versions, involves Achilles racing a tortoise. It has the near-immortal Achilles starting at point A, toward the same direction as a presumably slow tortoise starting at point B ahead of him. In his chase, Achilles first has to arrive at B. By then, the tortoise has advanced to B_1, so Achilles must next get to B_1. But by the time he gets to B_1, the tortoise has advanced to B_2. Achilles must then .... Consequently Achilles can never overtake the tortoise.

We can resolve the paradox by applying some numbers. Imagine that Achilles runs at 1000 cm/sec. (Can a human, even one with super powers, run that fast?) Assume the tortoise cruises at 1 cm/sec, and the original head start is 3000 cm. It takes Achilles 3 sec to cover the 3000 cm, during which time the tortoise advances 3 cm. Achilles needs another 3/1000 sec to cover the 3 cm, during which the tortoise zooms ahead another 3/1000 cm. Achilles must run another $(3/1000)/1000 = 3/1000^2$ sec, and so on. We now see what "never" amounts to. Achilles cannot overhaul the tortoise in 3 sec, nor

> in 3 + 3/1000, nor in 3 + 3/1000 + 3/1000^2, ... If the race clock does not reach
>
> $3 + 3/1000 + 3/1000^2 + ...$
>
> then Achilles does not reach the tortoise. But if the clock does make it to that mark (Exercise 3), then at that time Achilles catches up.

The relevance to Eudoxus's method is this: It does not suffice to note that doubling the number of sides *reduces* the area error. Imagine moving toward a chair that is initially 2 m ahead of you and has a wall 3 m beyond it. If in the first second you cover half the distance to the chair, the next second half the remaining distance, the next second half again, and so on, then it is true that you never reach the chair. You are getting closer to it: Your distance starts at 2 m, becomes 1 m, then ½ m, then ¼ m, ... You are also getting closer to the wall: Your distance from the wall starts at 5 m, becomes 4, then 3½, 3¼, and so on. The distinction is that your distance from the chair is decreasing *toward zero*. Distance from the wall is decreasing toward 3 m. In modern terms, we say that the areas of the polygons **approach** the area of the circle, meaning that the difference between the area of the circle and the areas of the polygons shrinks toward zero. For that, it is essential that the difference decrease by at least a fixed fraction (it does not have to be half) with each doubling of the number of sides. Mindful of Zeno's warnings about infinite processes, Eudoxus would have said that the areas of the polygons (necessarily smaller than the area of the circle) become greater than any fixed number smaller than the area of the circle.

Exercises III.A.4

1. Try a little exhaustion:
   a) Sketch (roughly) a *unit* circle and its inscribed regular hexagon. Use the hexagon's constituent triangles to compute its area exactly (in terms of radicals). Show that the result is approximately 2.60.
   (You may use a calculator. Of course Eudoxus could not, but he would have been able to do Exercise II.B.3:5 to approximate $\sqrt{3} \approx 97/56$ and calculate from there.)
   b) Sketch the inscribed dodecagon. Use its constituent triangles to compute its area in terms of *trigonometric* functions, and evaluate via a scientific calculator.
   (Here it is essential to emphasize that Eudoxus would not have used our trigonometric functions. We will see later how the Greeks did trigonometry.)
   c) If your answer in (b) is 3, then it is right and it should surprise you. Now explain why it works out to a whole number.
   d) Confirm that the answer in (b) and (c) eliminates more than half the error from the hexagon's approximation 2.60 to the circle's area π.
   e) Calculate the area as in (b) for the 24-gon, and do the error comparison as in (d).
2. Since the circumference of the unit circle is 2π, we could also approximate π by approximating half the circumference with half the polygon perimeters.
   a) What is the **semiperimeter** of the inscribed regular hexagon?
   b) Show that the semiperimeter of the dodecagon is 12 sin 15°, and calculate that.
   c) (Calculus) For the regular *n*-gon, the semiperimeter will be $n \sin(360°/2n)$. Find the limit of this expression as *n* tends to infinity. (Remember: The important limit relations for angles require the angles to be in radians.)
3. a) Find the "sum"
   $3 + 3/1000 + 3/1000^2 + ...$
   b) With Achilles trying to erase a 3000 cm gap at a closing speed of (1000 – 1) cm/sec, [I would have guessed that] it should take him 3000/999 sec. Does (a) agree?

5. Euclid

In the latter half of the fourth century, Greece was transformed. Before his death in 323 BCE, Alexander conquered much of the known world. Greek culture, including Greek mathematics, came to influence lands as far away as India.

Upon conquering Egypt around 330, Alexander established a new capital city in the Nile delta. (The Egyptians' capitals, like Memphis and Thebes, had been considerably up—south along—the river.) It was Alexander's habit to name his new cities "Alexandria." He also habitually left one of his Macedonian generals in charge of conquered lands. In the Egyptian Alexandria, it was Ptolemy. The Ptolemies, a dynasty of enlightened rulers, ruled Egypt for almost three centuries.

Ptolemy founded a center of the arts and sciences dedicated to the muses, called therefore the "Museum." It attracted scholars from throughout the Mediterranean world, and turned Alexandria into an important cultural center. Then he or his successors hit upon the idea of establishing a port tax. Any ship entering the port had to surrender its books, maps, and other sources of information long enough for them to be copied. The originals went to a collection in the Museum; the locals returned the copies. Thus was born the ancient world's most awe-inspiring institution: the Great Library. With this collection, Alexandria became the jewel of Mediterranean knowledge.

Euclid, already well known, was invited to this place around 320. His writings are a compendium of practically all mathematics and science known at the time. We will focus on his most famous book, *The Elements*. (Many leaders of schools gave their books that title.) It covers his geometry plus other material, some of which we will return to. It is important to note, right at the outset, that Euclid did not create or discover geometry; he *codified*, or organized, it.

a) geometry as a deductive system

The specific organization he used began with Eudoxus: that of a deductive system. In a **deductive system**, you have a body of knowledge divided into two kinds of statements: "axioms," which are assumed to be true; and "theorems," whose truth may be *inferred* from a combination of axioms and earlier theorems. [I will not make Aristotle's distinction between axioms and postulates; see **Boyer** for the ancient distinction.]

Imagine that we consider these four statements to be facts:

> a) If the three sides of one triangle are congruent respectively to the three sides of a second, then the two triangles are congruent.
> b) In a triangle, the angles sum up to a straight angle.
> c) If two sides of one triangle are congruent, then the opposing angles are congruent.
> d) In an isosceles right triangle, the acute angles measure 45°.

We do not need to *assume* that each is true. According to our argument in section III.A.1, the truth of the first guarantees the truth of the third. We may put (a) into the assumption group, as a payment that buys us the truth of (c). Once (c) is known, then (b) and (c) together allow us to conclude that in an isosceles right triangle, the two acute angles share 180° – 90° equally. From that, we deduce (d). Thus, putting (a) and (b) among our axioms *yields* (c) and (d) by deduction.

Once you identify some of your facts as theorems, they stay theorems. Not so with axioms. You examine your list of axioms. Suppose that on further investigation, you find that one of them follows from the others. Then you can move that one down to the top of the theorems list. You now have a shorter list of axioms, and all the theorems below follow from statements (axioms or theorems) above them. This process is clearly repeatable. You examine the shorter list of axioms, and if any of those follows from the rest, then you move it to the theorems list, and so on. The process cannot go on indefinitely, because presumably you begin with only a finite number of axioms. In fact, the process cannot even exhaust the axioms list, because you cannot draw information from *no* assumptions. (Actually, the trouble is that you can conclude *everything* from no assumptions, but that is a technical issue.)

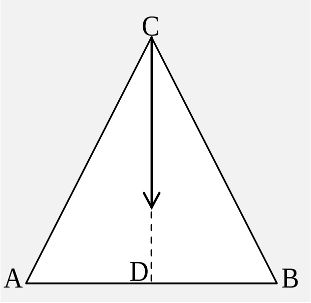
Clearly, then, the ideal basis for a deductive system is a set of axioms with three properties:
a) All the facts in your body of knowledge follow from the set.
b) None of the axioms follows from the others.
c) No contradiction follows from the axioms.

b) the view of Euclidean geometry

We have already made use of knowledge and methods from Euclidean geometry, and we will use plenty more, so we need not exhibit axioms and theorems of *The Elements* right here. For now, it is worth making two remarks.

One is that Euclid's organization of geometry was held to be the model of deductive reasoning for more than two thousand years. Indeed, *The Elements* was studied as much to provide grounding in deductive thinking as to give a body of geometric knowledge. This was true even though there were objections, right from the beginning, that Euclid's methods employed unstated assumptions. To illustrate what that means, consider this revisit to the base-angles theorem.

> Given triangle ABC (figure at right) with AC and BC congruent, draw the bisector (arrow in the figure) of angle C. Let it meet AB at D. Triangles ACD and BCD are congruent by SAS. Therefore angle A is congruent to angle B.



The argument is convincing, but it relies on an assumption that does not have justification by resort to axioms and theorems in Euclidean geometry. The assumption is that the bisector of angle C meets the segment AB. Certainly the picture suggests that the bisector must escape the triangle, and has to do that by crossing the bottom. This is frequently the situation in which the "hidden assumptions" come up: a relationship that is evidenced by a diagram but not by strict reliance on the deductive system. It happens that Euclid's axioms support the conclusion that the bisector must meet the *line* AB, but not that the meeting must be between A and B. Indeed, "between" is not a defined concept.

The other remark is about the different approaches to geometry and algebra. By the time of Euclid, and perhaps back to the time of Thales, the Greeks focused on geometry as a deductive pursuit. **Struik** names this focus the "Greek tradition." In what we can fairly call algebra, the Babylonians and Egyptians had a completely algorithmic (or problem-solving or methods-oriented) view. **Struik** refers to their way as the "Oriental tradition." The difference is not just of historical interest. We still maintain that separation in the way we teach algebra and geometry. We teach geometry as a deductive system, but algebra as a collection of methods.

6. Archimedes

Three giants led Greek mathematics during the fourth and third centuries BCE. We have already met Eudoxus, and will come back to him. We will soon get to know Apollonius. The greatest of them all was Archimedes the Syracusan.

Archimedes (284-212) was beyond famous; he was legendary. The ingenuity of his methods and clarity of his exposition made some swear he could not be human. His fame was doubtless known to Marcellus even before Marcellus's fleet had to contend with the inventions of Archimedes. When the Romans began to build an empire, they came into conflict with Carthage. (Where is Carthage?) During the second Carthaginian ("Punic") war, they decided to take Carthage's ally, the Greek city of Syracuse.

(Where is Syracuse?) Accordingly, Marcellus laid siege to the city in 214. His force encountered a bewildering array of defenses. They included road sections that sank or rose beneath marching units, cranes that reached out to lift and overturn boats, even (legend has it) an arc of men holding polished shields into the form of a concave mirror to focus solar light onto ships and set them on fire, all created by Archimedes. Marcellus gave strict orders that the engineer—back then, an "engine" was a weapon—was to be brought to him unharmed. The story goes that when the siege succeeded, one of the victorious Romans encountered the geometer doodling in the sand. Some say the old guy told the soldier to get out of his light, some that he simply failed to obey the soldier's order. Either way, in 212 the greatest mind of antiquity was stilled.

His inventions included the water screw, a helical device that could lift water to height limited only by the height of the screw. His exposition included the first explanation of the principle of the lever (equal products of force and lever arm) and, of course, of the hydrostatic principle that bears his name.

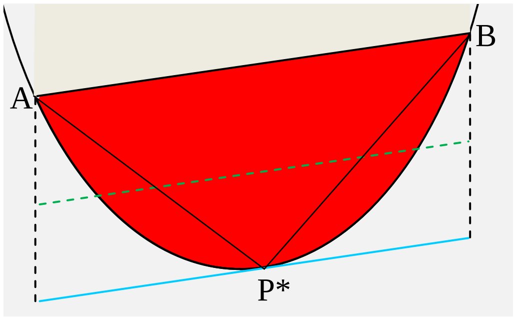
> **Archimedes's principle** states that a liquid exerts on an object a buoyant force equal to the weight of that liquid that has been **displaced** by the submerged part of the object. Set a 60 lb boat onto still water. It will settle down to where the submerged part of the boat, which includes a lot of air, takes the place of 60 lb of water. (How much volume would that be?) If you step into the boat, then to hold you and the boat afloat, the buoyant force has to increase by your weight. That is, the boat must settle further down to increase the displacement by whatever volume of water has your weight. Have eleven of your friends join you in the boat, and even with the entire boat under water, the buoyant force will not be sufficient to keep the dozen partiers dry.

a) ***Quadrature of the Parabola***

In the book of that title, his geometric work included the second squaring of a region with curved boundary.

(i) the area

The figure at right shows part of a parabola (black curve) having vertical axis and opening upward. We pick two points A and B on the parabola; they are shown, but need not be, on opposite sides of the axis. We refer to line segment AB as a **chord** of the parabola, and the region (red) bounded by the chord and the (**intercepted**) arc between A and B as a **segment** of the parabola.



Hold a ruler along AB, then slide the ruler downward without changing its inclination—in other words, so that its edge is always parallel to AB. As it slides through positions like the dashed green line, the edge continues to intersect the parabola at two points, but the points get closer together. It is intuitively clear that eventually, the two intersections will coalesce into one point P*. At that point, the edge traces the tangent (blue line) to the parabola, and the tangent is parallel to the chord.

Since all the points save P* of the parabola are above the tangent, triangle AP*B is the tallest one having AB as one side and the remaining vertex on the intercepted arc. Consequently AP*B has the greatest area of all such triangles. We distinguish it by naming it *the* **inscribed triangle**.

Archimedes squared the segment of the parabola with the following result.

Theorem 1. The area of the segment is 4/3 the area of the inscribed triangle.

(ii) the method of *The Method*

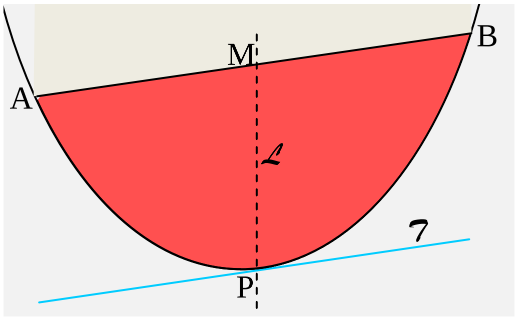
That “intuitively clear” four paragraphs back was not a guess at how the Syracusan thought. It was typical of Archimedes to discover mathematical truths informally. In the last figure, notice the vertical line segments (dashed black) from A and B to the tangent. They are opposite sides of a parallelogram that necessarily has twice the area of the inscribed triangle. (Why?) Archimedes used—of all things—a *weighing* argument. In effect, the explainer of the lever balanced the verticals within the (parabola’s) segment against those within the parallelogram on a lever with arms 3/2 as long on the segment side. (Observe the imprecise treatment of parallelogram and segment as the sum of a bunch of verticals.) He thus convinced himself that the segment has 2/3 the weight, meaning area, of the parallelogram.

We know he operated that way from his own testimony. In the short book titled *The Method*, he wrote that he always used analogy, heuristics, even physical comparisons, to discover relationships. Later, of course, he resorted to geometry to deliver proof. We will do likewise next.

(It was believed that *The Method*, like so much of Greek writing, had disappeared from the face of the Earth. The world only knew about it from second-hand references to it. Instead, a copy turned up in 1906, two millennia after its author was gone. It was in a manuscript, itself a thousand years old, carrying some of his other results.)

(iii) the method of Eudoxus

Now forget the intuition and work the geometry.

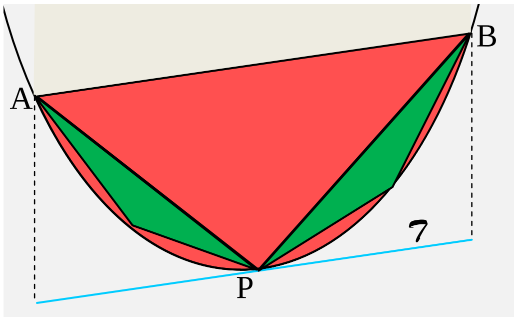


Start in the picture at right with the parabola and the chord AB. Add the midpoint M of AB and the (dashed) vertical line $\mathcal{L}$ through M. That line cuts the parabola at a point P.

Archimedes invoked a theorem of Apollonius. The theorem, which we [meaning you] will prove in the Apollonius section (Exercise III.A.7c:2a), says that all the chords parallel to AB have their midpoints along a single line parallel to the axis of the parabola. That line has to be $\mathcal{L}$.

Now add to the picture the (blue) line $\mathcal{T}$ through P parallel to AB. That line cannot have any other point of the parabola. If it did, P and the other point would bound a chord parallel to AB but with a misplaced midpoint. Accordingly, $\mathcal{T}$ is the tangent to the parabola at P. We have established that the tangent is parallel to the chord—showing us the inscribed triangle—at the point horizontally halfway between A and B. (Compare Exercise 6a.)

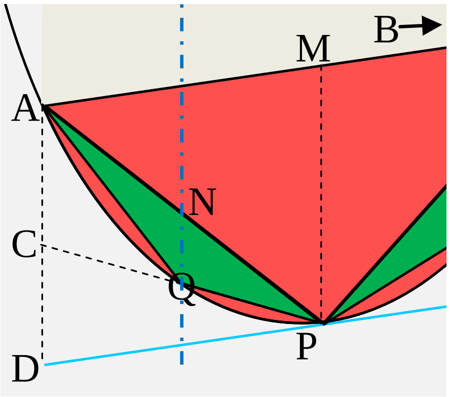
Let us nickname our segment the “level-0 segment” and ABP the “level-0 triangle.” In the figure at left, we bring back the verticals from A and B to the tangent $\mathcal{T}$ and draw in green the (“level-1”) triangles inscribed in the two (“level-1”) segments of the parabola outside the level-0 triangle. We noted that the level-0 triangle has half the area of the enclosing parallelogram. The parallelogram has greater area than the level-0 segment. Therefore the level-0 triangle covers more than half the area of the segment. For the same reason, each level-1 triangle covers more than half of its level-1 segment. At the next level, the four level-2 triangles cover more than half their level-2 segments. In approximation terms, each additional level eats away more than half the error in the area of the original segment. We conclude that the family of inscribed triangles exhausts the area of the original segment. (Compare Eudoxus’s exhaustion of the circle.)



(iv) measuring and adding the pieces

Finally, we measure the triangles by looking at the transition from one level to the next.

Focus on one level-1 segment. Recall what is true of every segment: The vertical (the parallel to the axis, dash-dot blue in the figure at right) through the midpoint N of the spanning chord AP is halfway between the verticals at A and P; and it crosses the parabola at the third vertex Q of that segment's inscribed triangle. Extend PQ to meet the vertical through A at C. Let D be where that vertical meets the tangent parallel to AB.



a) Q is the midpoint of CP. The reason is that equally-spaced parallel lines (the three verticals) cut any transversals into equal pieces. (See Exercise 3.)

b) The level-1 triangle AQP has half the area of ACP. That is because AQ is the median to side CP, and any median cuts its triangle into two triangles of equal areas (Exercise 4).

c) Triangle ACP has half the area of ADP. The reason is another contribution from Apollonius: C is the midpoint of AD (Exercise III.A.7c:2b). That means PC is the median to AD.

d) Triangle ADP has the same area as APM. In fact, those two are congruent; they are halves of parallelogram ADPM.

e) Triangle APM has half the area of the level-0 triangle APB, because PM is the median to AB.

Multiply the fractions to conclude that each level-1 triangle has 1/8 the area of the level-0 triangle.

If now we let T represent the area of APB, then the 2^1 level-1 triangles add

$$2^1T/8^1 = T/4$$

to the accumulating area. In the same way, each level-2 triangle has

1/8 the area of the adjacent level-1 triangle = 1/64 the area of APB.

Hence the 2^2 level-2 triangles together eat away another

$$2^2T/8^2 = T/4^2$$

from the segment's area. You see the pattern: The 2^n inscribed triangles of level-n together cover

$$2^nT/8^n = T/4^n$$

more area. Comfortable as we are with infinite processes, we conclude that the area of the segment is

$$\begin{aligned} A &= T + T/4 + T/4^2 + \dots \\ &= T\,(1/[1-1/4]) = 4T/3. \end{aligned}$$

It became Archimedes's habit to avoid Zeno's censure by using *finite* processes to bracket the value under investigation.

On one side, remember that the triangles of levels 0 through n do not completely cover the original segment; their areas add up to the underestimate

$$\begin{aligned} e_n &= T + T/4 + T/4^2 + \dots + T/4^n \\ &= T\,(1-1/4^{n+1})/(1-1/4) \qquad \text{(geometric sum formula, or factor the numerator)} \\ &= T\,(4/3-[1/3]/4^n). \end{aligned}$$

If you name any number smaller than $4T/3$—no matter how close—a big enough m will squeeze e_m between the number and $4T/3$. [Try an example: How big does m have to be to make

$(4/3-[1/3]/4^m)$ [no need for T] exceed $1.333\,333\,333 = 4/3-[1/3]/10^9$?]

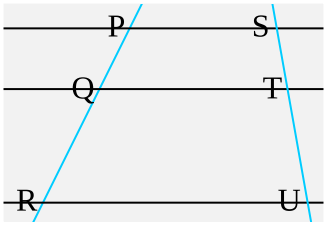
Since every e_m is less than A, Archimedes concluded that every number smaller than $4T/3$ is also smaller than A. That means $A < 4T/3$ is not possible.

> On the other side, recall that the parallelograms enclosing the level-n triangles *more than enclose* the segments left uncovered by the triangles of levels 0 to $n - 1$, and have twice the areas of the level-n triangles. Accordingly, if we add the triangle areas of levels 0 to $n - 1$ and the parallelogram areas of level n, we get the overestimate
>
> $$\begin{aligned} E_n &= T + T/4 + T/4^2 + \ldots + T/4^{n-1} + 2\, T/4^n \\ &= e_n + T/4^n \\ &= T\,(4/3 + [2/3]/4^n). \end{aligned}$$
>
> For the mirror-image argument, pick any number bigger than $4T/3$. Then any big enough k will squeeze E_k between $4T/3$ and the number. With every E_k exceeding A, Archimedes inferred that any number exceeding $4T/3$ also exceeds A. That renders $A > 4T/3$ impossible.

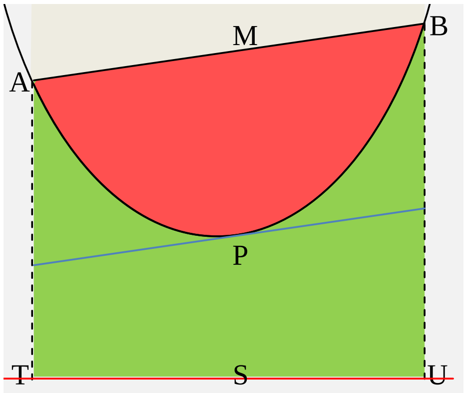
Only one possibility remains: The area of the segment has to be $4T/3$.

Exercises III.A.6a

1. To complete squaring the segment of the parabola: How would you construct a square with area 4/3 the area of a given triangle?
2. Given a parabola and one of its chords, how would you construct the tangent parallel to the chord? (Hint: You don't know *where* the axis is—not even whether it is vertical--but you do know that the place of tangency is on the "line of midpoints.")
3. Show that parallel lines cut transversals into proportional pieces. In detail: In the figure at right, the black lines are parallels, not necessarily equally spaced. The blue lines are transversals that may or may not intersect. One transversal cuts the parallels at P, Q, R; the other cuts at S, T, U. Show that
   PQ/ST = QR/TU = PR/SU.



4. Show that a median cuts its triangle into two triangles of equal areas.
5. At right, the parabola has vertical axis; chord AB has midpoint M and bounds the segment in red; the parallel tangent (blue) touches at P. Add a horizontal line (red) under the segment. The verticals (dashed) at A and B meet this line at T and U. The midpoint of TU is S. Show that the "area under the parabola" (the area A^* of the region filled in green) is given by
   $A^* = (1/6\ \text{AT} + 4/6\ \text{PS} + 1/6\ \text{BU})\ \text{TU}$.
   (Hint: The region consists of trapezoid ATUB minus the segment; the latter is 2/3 the ("enclosing") parallelogram determined by AB and the tangent; and M, P, and S have to be on one vertical, with MS being the average of AT and BU.)



   [Assume the parallelogram is entirely above the line, as shown. The formula still happens to be valid if the assumption is false. It likewise prevails if the parabola opens downward, even though in that case, the region is a trapezoid *plus* the segment.)
   [View the formula as a statement that the area under the parabola is
   $A^* = (\text{average height}) \times \text{width}$.
   There, "average height" refers to the *weighted* average that counts the height PS in the middle four times as much as the heights AT and BU at the two sides. The formula underlies **Simpson's rule**. The rule uses arcs of parabolas to approximate curves, and provides therewith a method for approximating integrals.]

6. (Calculus) In the Cartesian plane, consider the parabola given by $y = x^2$.
(You can answer this question for the general case $y = ax^2 + bx + c$, but the extra details are unnecessary. We will see later that the simpler form can be fitted to any parabola.)
a) Let A(a, a^2) and B(b, b^2) be two points on the graph. Show using the calculus that the point P(p, p^2), where the tangent to the parabola is parallel to AB, is halfway left-to-right between A and B; in other words, show that $p = (a + b)/2$.
b) Assume $a < b$. Integrate to find the area under the arc of the parabola from A to B.
c) The (average height) × (width) formula in Exercise 5 would read
$A = (1/6\ a^2 + 4/6\ ([a + b]/2)^2 + 1/6\ b^2)\ (b - a)$.
(Remember: Each y-value is the square of its x-value.) Does that match the answer in (b)?

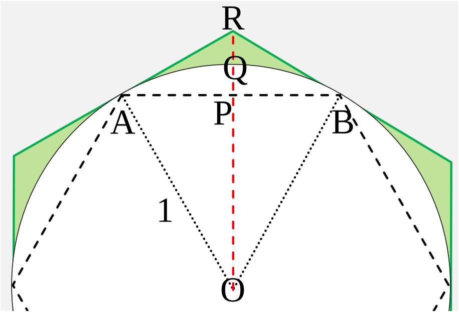
b) *Measurement of the Circle*

That title graced another book, the best-known work of Archimedes. In it, he brought two brilliant innovations to the method of exhaustion and its approximation of π.

(i) the overestimates

The first new idea was to use *circumscribed* polygons to produce overestimates bracketing π.

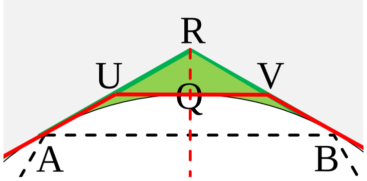
Start with the hexagons. At right, we have side AB of an inscribed regular hexagon (dashed outline, hereafter **inhexagon**) in the unit circle centered at O. You can build the circumscribed regular hexagon (hereafter **circumhexagon**) by constructing the tangents to the circle at the six vertices of the inhexagon. (How do you construct those tangents?) The tangents at A and B (dark green outline) meet at R. Each of RA and RB is half of a side of the circumhexagon (filled in light green). The line segment OR (dashed red) crosses the circle at Q and AB at P.



> AR and BR are congruent, because they are tangents from a common point. Hence triangles OBR and OAR are congruent by SSS, and OR bisects angle AOB. OR is therefore perpendicular to AB.
>
> Triangle OPA is a 30°-60°-90° triangle with hypotenuse OA = 1. It follows that its area is √3/8, and the inhexagon has area 12(√3/8) = 3√3/2. (Compare Exercise III.A.4:1a.). Triangle OAR is similar, but with longer leg OA =1. Hence its other leg is 1/√3, it has area 1(1/√3)/2, and the circumhexagon has area 12/(2√3) = 2√3. We bracket π between the two areas:
> (3/2)√3 < π < 2√3.

Having an overestimate to go with the underestimate is valuable, because we draw from them an indication of their accuracy. Write √3 ≈ 97/56 (Exercise II.B.3:5a). Our bounds become
291/112 < π < 388/112.
Those two differ by 97/112. Therefore their average is necessarily within 97/224 ≈ 0.43 of π; as it happens, the average is 679/224 ≈ 3.03.

Observe now that just as the inscribed (3×2^n)-gons exhaust the area of the circle, so the circumscribed ones squeeze the area. The circumdodecagon is tangent to the circle at the six vertices of the circumhexagon and at the midpoints of the vertices' arcs. At right, we magnify the top of the previous figure and add the tangent (red) at Q, meeting RA at U and RB at V. Here UV is one side, UA and VB half-sides, of the dodecagon (red outline). You can see that the circumhexagon overstates the area of the circle by six copies of the arrowhead shape



(filled green) bounded by RA, the arc from A to B, and BR. The dodecagon overstates by twelve copies of the smaller arrowhead bounded by UA, arc AQ, and QU. It is also clear that the dodecagon reduces the hexagon's overstatement by twelve copies of triangle UQR. Thereby the dodecagon cuts the hexagon's error *by more than half*.

> To explain: First, UV is closer to AB than to R. The reason is that UA = UQ, because they are tangents from the same point, whereas UQ < UR, because UR is the hypotenuse of right triangle UQR. The inequality UA < UR tells us that U is less than halfway up AR. Since UV is parallel to AB (Why?), we conclude that UV is less far above AB than it is below R.
>
> Now compare the shares (halves) of the errors leftward of RQ. The dodecagon's share (one of the smaller arrowheads) is smaller than triangle AQU. Triangle AQU, having the same base UQ as triangle UQR but smaller altitude (because UQ is closer to AB than to R), is smaller than triangle UQR. Those statements imply that the dodecagon's share constitutes less than half the hexagon's share. We conclude that the dodecagon's error is less than half the hexagon's. Compare Exercise 1.

The same reasoning applies to every doubling of the number of sides. We infer that the areas of the circumscribed (3×2^n)-gons approach π from above.

(ii) a subtlety

We wrote

$$291/112 < \pi < 388/112$$

after taking $\sqrt{3} \approx 97/56$. There is a caution we have to put in.

The estimate for $\sqrt{3}$ comes from Exercise II.B.3:5 on the square-root algorithm. The algorithm *always* yields overestimates. That means 388/112 overestimates $2\sqrt{3}$, which is the exact area of the circumhexagon, which overstates π. In other words,

$$\pi < \text{area of circumhexagon} = 2\sqrt{3} < 388/112.$$

Therefore the last is a legitimate overestimate for π. On the other side, we have

$$3\sqrt{3}/2 = \text{area of inhexagon} < \pi.$$

We may not immediately conclude

$$291/112 < \pi,$$

because 291/112 *exceeds* $3\sqrt{3}/2$.

We are about to produce arithmetic formulas—expressions using the four operations plus square roots—representing exact values for over- and underestimates of π. For those symbolic overestimates, we (without calculators) may apply the square-root algorithm to give slightly excessive numerical values. To give slightly low numerical values to the symbolic underestimates, we need to do extra work, as Archimedes must have done. For illustration, do Exercise 2.

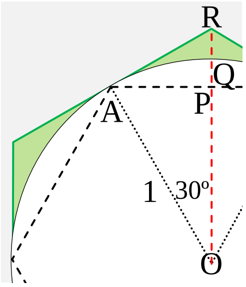
(iii) the calculations

It would be one thing to determine geometrically the inscribed and circumscribed values separately. The second innovation of Archimedes was to *interweave* them, turning the geometric question into one of sheer *arithmetic*.

We will modify the method of Archimedes. His calculations estimated *perimeters*, not areas. Thus, he would have found the inhexagon's perimeter to be 6 and the circumhexagon's to be $12(1/\sqrt{3})$, giving

$$6 < 2\pi < 4\sqrt{3}.$$

Perimeters offered him an advantage that will become obvious when we describe Greek trigonometry. Similarly seeking advantage, we will work in terms of our trigonometry. However, we continue to work with *areas*, since we have *proof* that the polygon areas approach the area of the circle. (See Exercise 5 for the perimeter computations corresponding to our area work.)



Revisit at right the earlier figure with the hexagons, but use trigonometry. In the inhexagon, the apothegm OP has length 1(cos 30°) and the half-side AP has 1(sin 30°). Therefore triangle OPA has area (cos 30° sin 30°)/2, and the inhexagon has 12 times that area,

$A_{6i} = 6 \sin 30° \cos 30°.$

In the circumhexagon, triangle OAR has one leg OA of length 1, the other AR of length 1(*tan* 30°). Triangle ORA has area (*tan* 30°)/2, the circumhexagon

$A_{6c} = 6\, tan\, 30°.$

Precisely the same reasoning applies to the dodecagon, with the number of sides doubled and the 30° angle cut in half. Thus,

$A_{12i} = 12 \sin 15° \cos 15°$ and $A_{12c} = 12\, tan\, 15°.$

(See Exercise 3. What would the formula be for 24 sides, 48, ...?)

Now observe that

$$A_{6i}\, A_{6c} = 36\, sin^2\, 30° \quad \text{(Justify!)}$$
$$= (A_{12i})^2.$$

That allows us to write

$A_{12i} = \surd(A_{6i}\, A_{6c}).$

It allows us to calculate A_{12i} in terms of radicals, as both A_{6i} and A_{6c} are: We know

$A_{6i} = 3\surd 3/2$ and $A_{6c} = 2\surd 3.$

In other words, it gets us from A_{6i} and A_{6c} to A_{12i} via arithmetic.

Separately, with a similarly desirable result but a lot more work, we find

$$\begin{aligned} 1/2\ (1/A_{6c} + 1/A_{12i}) &= 1/(12\, tan\, 30°) + 1/(24 \sin 15° \cos 15°) \\ &= (cos\, 30° + 1)/(24 \sin 15° \cos 15°) \\ &= \cos 15°/(12 \sin 15°) \\ &= 1/A_{12c}\,. \quad \text{(Justify all four equalities.)} \end{aligned}$$

We get from A_{6c} and A_{12i} to A_{12c} arithmetically.

Note that the key to the method is halving the central angle. Consequently the method is recursive; it reapplies every time you double the number of sides. Thus,

$A_{24i} = \surd(A_{12i}\, A_{12c})$ and $A_{24c} = 1/[1/2\ (1/A_{12c} + 1/A_{24i})], \ldots.$

Archimedes carried out the perimeter/circumference approximations up to the 96-gon. His estimates are rendered most simply as

$$3 + 10/71 < \pi < 3 + 10/70.$$

(See **Struik**, page 53, for elaboration.) Note that the two numbers differ by 10/4970 ≈ .002. Necessarily one of their estimates of π is accurate to three decimal places. (It happens that 3+10/71 is closer. Its error is actually only about .0008, less than $1/2^4$ times the error or the hexagon. See also Exercise 4.)

The expression $1/2\ (1/A_{12c} + 1/A_{24i})$ and its reciprocal did not materialize out of thin air. Given positive numbers *u* and *v*, the expression $1/[1/2\ (1/u + 1/v)]$ is called their **harmonic mean**. (Render it in words: It is the reciprocal of their average reciprocal.) It is one of many averages the Greeks studied, and the name reflects its origin in their study of music. The square root $\surd(uv)$ of their product is another average, called the **geometric mean** (or **mean proportional**) of *u* and *v*. That name reflects the fact that *u*, $\surd(uv)$, and *v* are in geometric progression; see Exercise 6.

Exercises III.A.6b

1. We saw that the unit circle's circumhexagon has area $2\sqrt{3}$ and the circumdodecagon has area $12 \, \textit{tan} \, 15°$ (from 3c below). Show that the latter's error is less than half the former's.
2. a) Describe a method to find a rational *underestimate* for $\sqrt{3}$.
   b) Execute the method, then check that your approximation really is just under $\sqrt{3}$.
3. a) Show that the area of the indodecagon of a unit circle is $12 \sin 15° \cos 15°$.
   b) Write the exact values of sine and cosine of 15° (in terms of radicals), then calculate the answer to (a). (Compare Exercise III.A.4:1b and c.)
   c) Show that the area of the circumdodecagon is $12 \, \textit{tan} \, 15°$.
   d) Show that in general, the circumscribed regular n-gon has area ($n \, \textit{tan} \, [360°/2n]$).
   e) (Calculus) Find the limit of the expression in (d) as n tends to infinity.
4. Use a spreadsheet or programmable calculator to determine which fractions are closer to π than 22/7 is, of the fractions between 3 and 4 that have denominator from 2 to 100.
   (You have to look at fractions $3+i/j$, where $2 \le j \le 100$. You need not avoid duplication; you can look at both 3+1/10 and 3+2/20. But since you know that $\pi < 3+15/100$, save time by limiting i to 1 through 15, or better, to 1 through $j/6$. The results will show you why we use 22/7 for hand calculation. You will also find that 3+14/100 is not on the list.)
5. a) Show that the perimeter P_{6c} of the circumhexagon to a unit circle is $12 \tan 30°$.
   b) The perimeter P_{6i} of the inscribed hexagon is $P_{6i} = 6 = 12 \sin 30°$. Show that the circumscribed and inscribed dodecagons have
   $P_{12c} = 1/(1/2\,[1/P_{6c} + 1/P_{6i}])$ (the harmonic mean),
   $P_{12i} = \sqrt{(P_{6i}\, P_{12c})}$ (the geometric mean).
   Notice that with perimeters, you calculate the circumscribed before the inscribed.
   c) Show that in general, the perimeter of the circumscribed n-gon is $2n \tan (360°/2n)$.
   d) (Calculus) Find the limit of (c) as n tends to infinity. Why is it not π?
6. Given positive numbers u and v, let $g = \sqrt{(uv)}$ be their geometric mean, $h = 2uv/(u + v)$ their (simplified) harmonic mean, $a = (u + v)/2$ their **arithmetic mean**. Show that:
   a) u, g, and v are in geometric progression. As a result, g is between u and v.
   b) $g \le a$. That explains Exercise II.B.3:5d.
   c) g is the geometric mean of h and a. That implies h, g, and a are in geometric progression, and forces $h \le g \le a$.
   d) g is the same fraction of u that v is of g.
   e) h is fractionally below u as far as it is fractionally above v; that is,
   $(u - h)/u = (h - v)/v$.

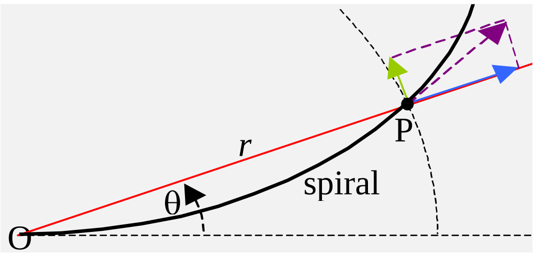
c) foreshadowing the calculus

At the introduction of integrals, our calculus courses say that in squeezing the circle between inscribed and circumscribed polygons, Archimedes anticipated the methods of integral calculus. In fact, he more directly anticipated methods in both differential and integral calculus, in studying the question of tangent to a curve other than a circle and the technique we now call summation of parts.

(i) tangent to a curve

The Greeks had two ways to describe curves. One was as the intersection of surfaces, the other as the trace of a moving point. We can describe the parabola, for example, either way: It is the intersection of a certain cone and plane, and it is the path traced by a point so moving as to remain equidistant from a certain point and line. The **spiral of Archimedes** is traced by a point starting at the end of a half-line and moving uniformly away from the end while the half-line, its end fixed, rotates at a uniform rate.

In the figure at right, the red line starts at the horizontal position (dashed black line) and rotates counterclockwise with its endpoint O fixed. As it rotates, the point P moves outward along the line. Thus P traces out the spiral indicated by the heavy black curve. Archimedes described its tangent in these terms.



Let us say the rotation rate is 0.5 radian/sec and P moves out at 3 cm/sec. (Half a radian is about 28°; the line goes all the way around in about 13 sec.) If we set r = OP and denote the angle between the line and the horizontal by θ, then we see that

$$r = 3t \text{ and } \theta = 0.5t.$$

In the language of polar coordinates, the spiral has the familiar equation $r = 6\theta$.

The figure captures the instant $t = \pi/3 \approx 1.05$ sec, when $\theta = 30°$ and $r \approx 3.14$ cm. At that time, P is moving to the right and upward. If we specify its precise direction, we arrive at the direction of the tangent to the spiral.

Archimedes specified it by thinking of P as having a **radial** (outward) motion along the line and a **tangential** (circling) motion *perpendicular* to the line. The radial motion—what the language of physics calls the **radial component of velocity**—is indicated by the blue arrow. It has constant speed, the length of the arrow, of 3 cm/sec. The tangential motion, tangent to the circle (dashed black arc) centered at the origin and reaching P, is indicated by the green arrow. It does not have constant speed. Remember that 0.5(radian)/sec is an **angular speed**, the rate at which θ grows. The tangential **linear speed**, the rate at which distance is covered perpendicular to OP, is this angular speed multiplied by r; it grows as P moves out. At our particular instant, the green arrow has length (.5/sec)(3.14 cm) ≈ 1.57 cm/sec.

The combination of the two motions—the **resultant**, physics says—is the diagonal of the rectangle they determine, the rectangle completed by the dashed violet lines. That diagonal is the dashed violet arrow. Its length comes easily from the Pythagorean theorem. The length is the speed of P, which interests us no more than it did Archimedes. What we want is its direction. The rectangle makes clear that the angle between the violet arrow and the blue has (trigonometric) tangent $1.57/3 \approx 0.52$. Therefore at our instant, the direction of the tangent to the spiral is at angle $(30° + \arctan 0.52) \approx 58°$ above the horizontal. In general, its direction at time t is

$$\theta + \arctan\,(\text{tangential speed/radial speed}) = 0.5t + \arctan(1.5t/3)$$

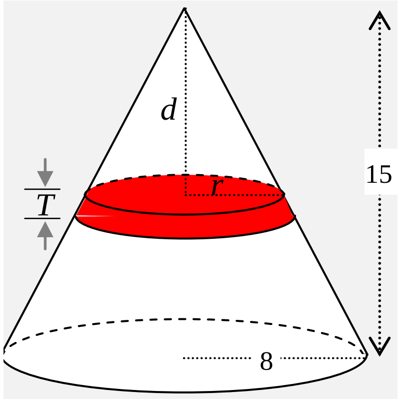
counterclockwise around from horizontal.

[This idea of analyzing motion in terms of how it covers distance in each of two perpendicular directions is a routine part of modern mechanics. As far as I know, it was not picked up until Galileo’s studies of falling objects, more than 1800 years after Archimedes.]

The spiral of Archimedes, assuming you could construct it, would answer the problem of trisecting an angle. In our figure, segment OP is at 30° above horizontal. Trisect the *segment*, putting point Q one-third from O to P. Draw the circle centered at O and reaching Q. It intersects the spiral at a point R where $r = \pi/3$ and $\theta = r/6 = \pi/18$. Therefore OR trisects the original 30° angle.

(ii) volume by summation

Archimedes used the technique of summation of parts, which underlies the development of integral calculus, to derive a number of volumes and areas. We will first illustrate with the formula for the volume of a cone. Archimedes attributed the formula to Democritus and said that Eudoxus had given a proof, presumably a geometric one.



In the figure at right, we see a right circular cone with vertical axis, height 15, radius 8. We cut through it with a large number, say 999, of uniformly spaced horizontal planes. These cut the cone into 1000 slices, all of them $T = 15/1000$ thick.

Imagine that the solid slice filled in red is #23 down from the vertex. All its horizontal cross-sections are circles. Its top face, at depth d below the vertex, has radius r. The top is the 22nd cut, so it has

$d = 22T$.

From similar triangles, we have

$r/d = 8/15$,

wherefore

$r = (8/15)d = (8/15)\,22T$.

That is the minimum radius for any horizontal cross-section in this slice; the circular sections expand as you go lower. Consequently the volume V_{23} of the slice exceeds the volume of a right circular *cylinder* of that radius and thickness. In symbols,

$V_{23} > \pi\,[\text{radius}]^2\,(\text{thickness}) = \pi\,[(8/15)\,22T\,]^2\,T$.

At the same time, the bottom face of the slice is the biggest horizontal cross-section of the slice. Its depth from the vertex is $23T$, making its radius $(8/15)\,23T$. The volume of the slice must be less than the volume of a cylinder with this radius and thickness T. That makes

$V_{23} < \pi\,[(8/15)\,23T\,]^2\,T$.

The same reasoning applies to all the slices. Their volumes are bracketed by

$\pi\,[(8/15)\,0T\,]^2\,T < V_1 < \pi\,[(8/15)\,1T\,]^2\,T,$

...,

$\pi\,[(8/15)\,22T\,]^2\,T < V_{23} < \pi\,[(8/15)\,23T\,]^2\,T,$

...,

$\pi\,[(8/15)\,999T\,]^2\,T < V_{1000} < \pi\,[(8/15)\,1000T\,]^2\,T.$

(To check: Notice that $1000T = 15$, which is how far the bottom of the last slice is from the vertex.)

Summing the parts, we bracket the volume V of the cone. We can add those numbers because they have a pattern. Rewrite

$\pi\,[(8/15)\,22T\,]^2\,T = \pi\,8^2/15^2\,22^2\,(15/1000)^3 = 22^2\,\pi\,8^2\,(15)\,/1000^3.$

The π and all the factors that follow are common to all the numbers. Factor them out, and the numbers in the last line add up to

$(0^2 + 1^2 + ... + 998^2 + 999^2)\,\pi\,8^2\,(15)\,/1000^3.$

We know what those squares add up to (Would Archimedes have known?):

$0^2 + 1^2 + ... + 998^2 + 999^2 = (999)(1000)(1999)/6.$

We conclude

$(\mathbf{.999})(\mathbf{1.999})/6\,[\pi\,8^2]\,(15) < V.$

In the same way, adding up the numbers on the right leaves us with

$V < (\mathbf{1.001})(\mathbf{2.001})/6\,[\pi\,8^2]\,(15).$

You can see that if we had made it a million slices, then the factors in bold would have been

$$(\mathbf{.999999})(\mathbf{1.999999}) \qquad \text{and} \qquad (\mathbf{1.000001})(\mathbf{2.000001}).$$

Archimedes concluded that if you name any number below

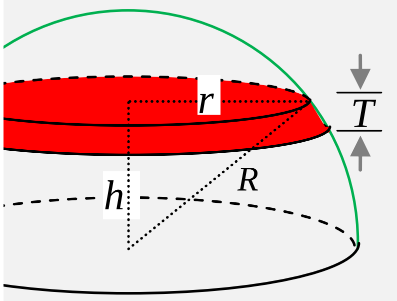
$$(\mathbf{1})(\mathbf{2})/6\ [\pi\ 8^2]\ (15)\ ,$$

then V exceeds that number. Similarly, if you name any number beyond $(\mathbf{1})(\mathbf{2})/6\ [\pi\ 8^2]\ (15)$, we can by slicing the cone into sufficiently many pieces show that the volume is smaller than that number. Only one conclusion is possible:

$$V \quad = \quad 2/6\ [\pi\ 8^2]\ (15) = \quad 1/3\ [\text{area of the base}]\ (\text{height}).$$

This (literal) analysis has an interesting consequence. Imagine *any* cone that has a height of 15 and base area $\pi(8)^2$, even if that base is some blob (or a polygon, as with a pyramid). Similarity would still make the linear dimensions of the faces (top and bottom) of the slices proportional to depth. That would make the areas of the faces proportional to $(\text{depth})^2$. Those slices would then have the same volumes as the slices of the right circular cone. We would have the same parts to sum, and we would conclude that the blob-based cone has the same volume as our right circular cone. The Greeks must have known the principle: If two solids of the same altitude have cross-sections of equal area at each horizontal level, then they must have equal volumes. Yet it was not until 1635 CE that Buonaventura Cavalieri, a student of Galileo, explicitly stated this "Cavalieri's principle."

The same kind of analysis gives us the volume of a sphere.



The figure at right shows about a third of the outline (dark green) of the sphere of radius R. We use 1999 equally spaced horizontal planes to cut the sphere into slices $T = 2R/2000$ thick.

Say the solid slice illustrated (red) is #23 upward from the equator. Its top face is a circle, at height

$$h = 23T$$

above the equator. The face has radius

$$r = \sqrt{}(R^2 - h^2).$$

It is the slice's smallest cross-section. Consequently the volume V_{23} of the slice has

$$V_{23} \quad > \quad \pi\, r^2\, T \quad = \quad \pi\ (R^2 - 23^2\, T^2)\ T.$$

Similarly, the bottom face has radius $\sqrt{}(R^2 - 22^2\, T^2)$. The slice's volume has

$$V_{23} \quad < \quad \pi\ (R^2 - 22^2\, T^2)\ T.$$

Reasoning likewise for the other slices above the equator, we write

$$\pi\ (R^2 - 1^2\, T^2)\ T \quad < \quad V_1 \quad < \pi\ (R^2 - 0^2\, T^2)\ T,$$

$$\ldots,$$

$$\pi\ (R^2 - 1000^2\, T^2)\ T \quad < \quad V_{1000} \quad < \quad \pi\ (R^2 - 999^2\, T^2)\ T.$$

Summing the volumes, evaluating the sums of the squares, and substituting $T = R/1000$, we find (Exercise 3) that the volume V of the upper hemisphere satisfies

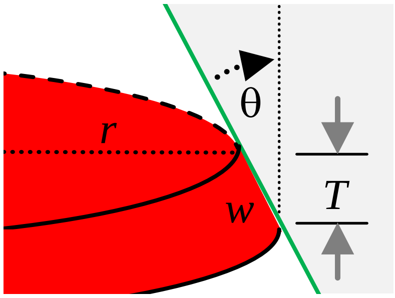
$$\pi\ (R^3 - 1.001[2.001]/6\ R^3) \quad < \quad V \quad < \quad \pi\ (R^3 - .999[1.999]/6\ R^3).$$

As with the cone, we are led to conclude that the hemisphere encloses a volume

$$V \quad = \quad \pi\ (R^3 - 2/6\ R^3) \quad = \quad 2/3\ \pi R^3.$$

(iii) surface area by summation

We can apply the same technique to determine the area of the curved part of the right circular cone.



Look at the edge of our slice #23 from the 8×15 cone, magnified at right. Denote by θ the angle between the **element** of the cone—the (green) line along the cone through the vertex—and the vertical. (The angle is not an independent quantity; θ is given by

$\tan\theta = 8/15$.)

The curved surface of the slice is a ribbon of uniform width

$w = T/\cos\theta$. (Why are w and T related that way?)

However, the ribbon has unequally long top and bottom edges. The top edge has length, meaning circumference,

$2\pi r \;=\; 2\pi\,(22T\,8/15)$.

The bottom edge has circumference

$2\pi\,(23T\,8/15)$.

Therefore the ribbon's area S_{23} satisfies

$$2\pi\,(22T\,8/15)\,T/\cos\theta \quad < \quad S_{23} \quad < \quad 2\pi\,(23T\,8/15)\,T/\cos\theta.$$

Summing the 1000 areas and remembering that

$T = 15/1000$, $\cos\theta = 15/\sqrt{(8^2 + 15^2)}$,

we find (Exercise 4a) the curved area of the cone to be

$S = \pi\, 8\, \sqrt{(8^2 + 15^2)}$.

It is worthwhile to attach interpretation to the formula. The length

$H = \sqrt{(8^2 + 15^2)}$

is called the **slant height** of the cone. Our formula reads

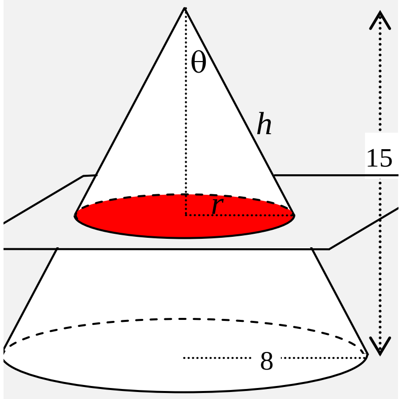
$S = \pi$ (radius) (slant height).

Think of the cone as a squeezed cylinder, with circumference $2\pi(8)$ at the bottom, reduced to zero at the top. That gives it average circumference

$(0 + 2\pi 8)/2 = \pi 8$.

Multiply average circumference by the "height" H of the "cylinder," and you get the cone's surface area.

We can give the same interpretation to the surface area of a certain frustum. A **frustum** of a right circular cone is the solid below any plane cutting the cone parallel to the base.



At right, the cutting plane's circular section (red) has radius r and leaves slant height h above. We know the curved area above the plane is πrh. That makes the curved area of the frustum

$S^* = 8\pi H - \pi rh$.

The section has circumference $2\pi r$. Therefore the average circumference of the frustum is

$(2\pi r + 2\pi 8)/2 = \pi\,(r + 8)$.

The product of this average circumference and the lower slant height $H - h$ matches S^* (Exercise 4b).

Exercises III.A.6c

1. Archimedes is sometimes described as the inventor of integral calculus. What parts of his work justify such description?
2. ("The Pirate Problem") A Coast Guard boat is stopped in a deep fog. For a second, the fog lifts, and the boat spots the pirate vessel it is looking for. The boat gets a precise fix on the

pirate's location, but immediately the fog comes back down. The Guard knows that the pirate, having seen the Coast Guard boat, will flee in a straight line at 20 knots, but not in what direction. The Guard boat can do 25 knots. Describe a path the Guard can follow to guarantee that it encounters the pirate vessel. (Hint: It is a spiral.)

3. Show that

$$\pi\,(R^2 - \mathbf{1}^2[R/1000]^2)\,[R/1000] + \ldots + \pi\,(R^2 - \mathbf{1000}^2[R/1000]^2)\,[R/1000]$$
$$= \pi\,(R^3 - 1.001[2.001]/6\,R^3).$$

4. a) With $T = 15/1000$ and $\cos\theta = 15/\sqrt{(8^2 + 15^2)}$, show that

$$2\pi\,(\mathbf{1}T\,8/15)\,T/\cos\theta + \ldots + 2\pi\,(\mathbf{1000}T\,8/15)\,T/\cos\theta \quad = \quad 1.001\,\pi\,8\,\sqrt{(8^2 + 15^2)},$$
$$2\pi\,(\mathbf{0}T\,8/15)\,T/\cos\theta + \ldots + 2\pi\,(\mathbf{999}T\,8/15)\,T/\cos\theta \quad = \quad 0.999\,\pi\,8\,\sqrt{(8^2 + 15^2)}.$$

b) In the figure of the frustum, show that the known surface area

$$S^* = 8\pi H - r\,\pi h$$

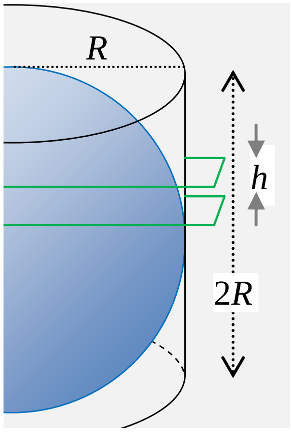
matches the product

$$\pi\,(r + 8)\,(H - h)$$

of average circumference and (lower) slant height. (Hint: Bring in θ.)

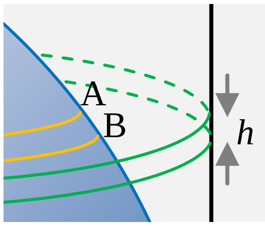
---

d) surface area of the sphere



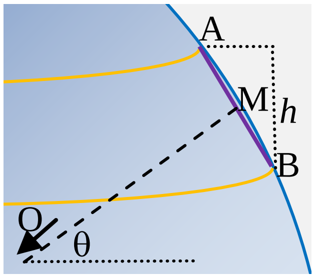
Archimedes used an imaginative comparison to find the area of the sphere.

Fit the sphere into an equally tall, equally wide right circular cylinder, as in the figure at right. We will say that the sphere is inscribed in the cylinder, the cylinder "circumscribed" about the sphere. Then cut through cylinder and sphere with two horizontal planes (green outlines) an tiny distance h apart.



From the cylinder (heavy black edge in the figure at left), the two planes cut a smaller cylinder. Around this smaller cylinder, the curved part is a ribbon (bordered by green) of uniform width h and constant circumference $2\pi R$ at every horizontal level. Consequently its area is simply $2\pi Rh$. [Fold a rectangular piece of paper into a right circular cylinder, to see that a horizontal band from it unfolds into a rectangle of dimensions circumference × height.]

In the same figure, the two planes cut a band (bordered by orange) from the sphere. In the plane of the page, the band has point A at the top, B at the bottom. Imagine that h is so small that the arc AB is indistinguishable from *line segment* AB, as suggested at right, where we magnify arc and segment (violet). Now the band is indistinguishable from the curved surface of a frustum. Make M the midpoint of AB. If the radius OM of the sphere makes an angle θ with the horizontal, then M is on a horizontal circle of radius ($R\cos\theta$). That circle's circumference is the average for the frustum. By the previous subsection (iii), the frustum has (curved) surface area



$$\text{(average circumference) (slant height)} = 2\pi(R\cos\theta)\,\text{AB}.$$

Because OM has to be perpendicular to segment AB, the latter makes an angle θ with the *vertical*. Hence

$$h/\text{AB} = \cos\theta,$$

and the area of the frustum is

$$(2\pi\,R\cos\theta)\,h/\cos\theta = 2\pi Rh.$$

The frustum—the band intercepted on the sphere—has the area of the band on the cylinder.

That is a surprising result. It implies that the surface area of the part of the sphere between two parallel cutting planes is proportional to the distance between them. For example, the region between the equator and latitude (same as θ) 30°—on Earth, that would span from the Galapagos Islands to New Orleans—which is half as tall as the northern hemisphere, has one-quarter the area of the sphere. It implies as well, of course, that the area of the sphere equals the (curved) area of the cylinder.

[They say that this discovery was the favorite of Archimedes, so much so that he requested the picture of the sphere in the cylinder be carved into his tombstone. Somebody complied, though the marker is long gone. My fond dream is that it was Marcellus.]

Exercises III.A.6d

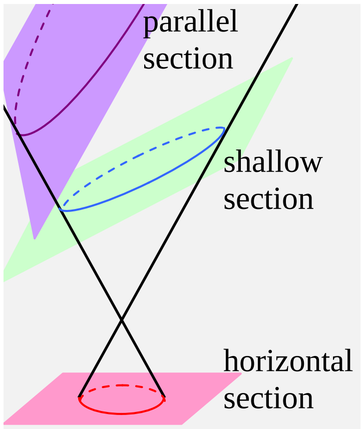
1. Archimedes described the volume within the sphere as two-thirds the volume within the circumscribed cylinder. Does that reflect our calculation in III.A.6c(ii)?
2. What is the ratio between the volume within the sphere and the volume enclosed by the right circular cone whose vertex is the north pole and whose base, tangent to the sphere at the south pole, has the radius of the sphere? (Hint: Draw the picture.)
3. Archimedes described the area of the sphere as four great circles. Does that agree with our statement that the sphere has the area of the circumscribed cylinder?

7. Apollonius

Apollonius was born in Perga, in modern-day Turkey. His birth and death followed those of Archimedes by about twenty years; their lives overlapped for fifty. He made extraordinary studies of curves. We will give an indication of the extent of his work by spending a long section on a miniscule part of it, in particular on the familiar curve called the ellipse.

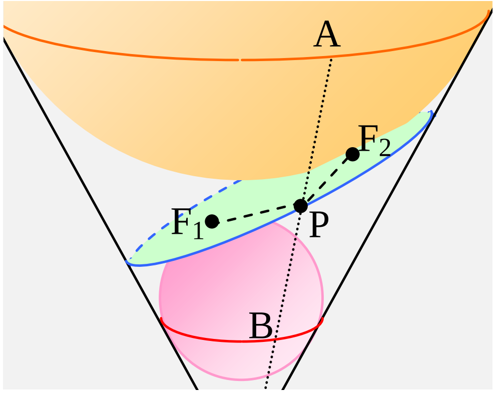
a) conic sections

Apollonius was the first to describe what we call **conics** in terms of sections of just one cone. In the figure at right, we see part of both halves ("naps") of a right circular cone (edged by the two black lines) with vertical axis. The cone is "right circular" because its **section** (intersection) by a horizontal plane is a circle (pink plane, red circle). If we incline the plane at a shallow angle, the section (green plane, blue section) becomes elongated left-to-right more than front-to-back. The resulting oval is an **ellipse**. The steeper the angle, the more oblong is the ellipse, until the plane's inclination matches that of the side of the cone. At that stage, the nature of the section changes suddenly (nowadays: "catastrophically"). The part of the curve on the near side of the cone (solid purple curve) diverges from the far part (dashed purple); the curve never closes. That section is a **parabola**. Beyond that inclination (not illustrated), the section remains an open curve, called a **hyperbola**. At first look, there is only one obvious difference between the hyperbola and parabola: The hyperbola has two separate **branches**, because the steep plane necessarily cuts the bottom half of the cone as well as the top. We will characterize the ellipse, then relate it to the parabola and hyperbola and see one difference in nature between the last two.



b) the ellipse

In the next figure, we add some objects to the image of the shallow section. Imagine resting a small spherical balloon (pink) on the inside of the cone's surface, below the shallow plane. Inflate it until it grows just big enough to touch the plane at a point F_1. The balloon, then, is tangent to the plane at F_1 and tangent to the cone along a circle of latitude (red) below the balloon's equator. Next, rest a giant light-orange spherical balloon on the inside of the cone, above the plane. Deflate that one until it sinks just low enough to touch the plane at F_2. This sphere is tangent to the plane at F_2 and tangent to the cone along the orange circle. We pick an arbitrary point P of the ellipse. The figure shows the segments (dashed black) PF_1 and PF_2. Finally the figure has part of the element of the cone, the line joining P and the vertex of the cone. That part (dotted black) meets the orange circle at A and the red one at B. We will use all of those to show that what is special about the ellipse is that the distances from any P to F_1 and F_2 have a constant sum.



(i) the distance characterization

Each of F_1 and F_2 is a **focus** of the ellipse. The segments PF_1 and PF_2 are the **focal radii** from P. Because the cutting plane and the cone are tangent to the upper sphere, PA and PF_2 are tangent to that sphere. In plane geometry, the two tangents to a circle from a given point outside the circle have to be equally long. Likewise in three dimensions: Any of the infinity of tangents to a sphere from one point outside the sphere are equal. Thus,

$PF_2 = PA$.

Similarly, because PB and PF_1 are tangent to the lower sphere,

$PF_1 = PB$.

Therefore

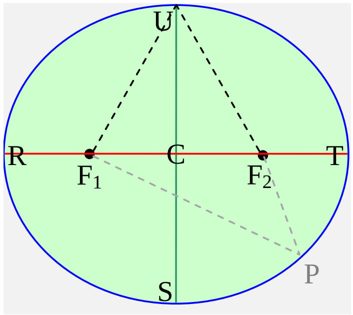
$PF_1 + PF_2 \quad = \quad PB + PA \quad = \quad AB$.

Now consider: No matter where on the ellipse P is, the length AB is the same. It is the distance *along the cone* between the horizontal plane that contains the orange circle and the one that holds the red circle. This is the tangent principle at work again. Let V be the cone's vertex, located below the figure. The tangents from V to the orange sphere form the cone. No matter where on the orange circle A is, the length VA is the same. Similarly, VB has a fixed length independent of where B is on the red circle. Therefore AB = VA – VB has constant length.

We have found that for the points on the ellipse, *the sum of the focal radii is constant*. The ellipse is the plane locus of points whose distances from two certain fixed points add up to a certain fixed number.

[We will begin to identify curves as **loci** (plural of **locus**, Latin for "place"). View Exercise 4.]

Keeping that characterization in mind, forget the cone for a bit and look face-on at the ellipse. In this figure, we have the foci F_1 and F_2, plus the typical point P and its focal radii (grayed). We draw the line (red) joining the foci, meeting the ellipse at R and T. The ellipse is symmetric about that line, because each point in the upper half has the same (length) focal radii as the point below it in the lower half. Then we draw the perpendicular bisector (green) of F_1F_2, meeting the ellipse at S and U. The ellipse is symmetric about that line as well, because a point in the left half has the same focal radii, but in opposite order, as its brother in the right. We call

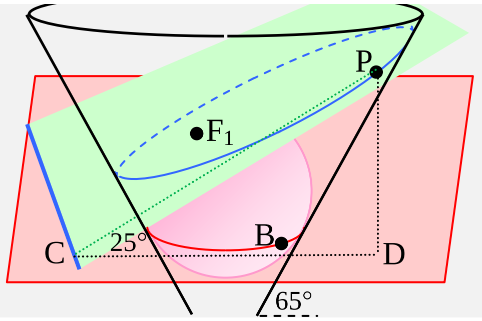


each line an **axis of symmetry** of the ellipse. Their intersection C is the **center of symmetry**, or simply **center**. The figure also has the focal radii from U, and is faithful to the appearance of the ellipse.

> Write 2*s* for the constant sum of the focal radii, 2*m* for the height US, and *f* for the **focal distance** $F_1C = CF_2$. Since R and U are on the ellipse, each one's focal radii must sum to 2*s*. For R, we have
>
> $$RF_1 + RF_2 = 2s.$$
>
> By symmetry, $RF_2 = F_1T$. Therefore
>
> $$\begin{aligned} 2s &= RF_1 + RF_2 \\ &= RF_1 + F_1T \\ &= RT. \end{aligned}$$
>
> In words, the width of the ellipse is the constant sum of the focal radii. For U, we have
>
> $$UF_1 + UF_2 = 2s.$$
>
> But UF_1 and UF_2 are equal, because US is the perpendicular bisector. Therefore
>
> $$UF_1 = UF_2 = s.$$
>
> We now see that UCF_1 is a right triangle with legs $UC = m$ and $CF_1 = f$, hypotenuse *s*. Consequently
>
> $$m^2 + f^2 = s^2.$$
>
> Necessarily $s > m$; the axis with the foci is longer than the other one. Hence we call RT the **major axis**, R and T the **major vertices**, US the **minor axis**, U and S the **minor vertices**, of the ellipse.

(ii) the focus-directrix characterization

Now we go back to the ellipse on the cone.



The figure here loses everything related to the upper balloon, puts back the sectioning plane (green), and adds the horizontal plane of the red circle (pink). That plane is entirely below the ellipse; any P is as far above the plane, along its element of the cone, as it is far from F_1. The red plane and the sectioning plane intersect along the solid blue line. For simplicity, give the inclination of the (element of the) cone a specific value, 65°, and the inclination of the cutting plane 25°.

Draw the perpendicular (dotted green) from P to this line, meeting the line at C. PC then has the inclination of the cutting plane, which we have set at 25°. Add the vertical through P, intersecting the horizontal plane at D. Notice that triangle PDC is in a plane parallel to (and in front of) the plane of the drawing. On the other hand, B (along the element of the cone through P) is behind the plane of PDC (owing to the tapering of the cone), so that PDB's plane points into the plane of the drawing. The picture will give us a second characterization of the ellipse.

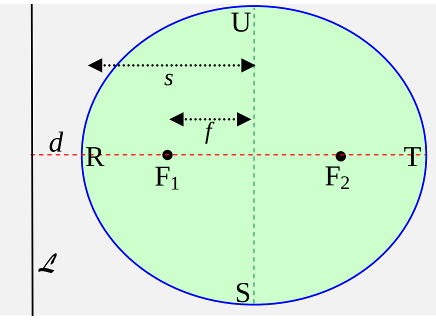
> In the plane that has P, D, and C, right triangle PDC has a 25° angle at C. Therefore
>
> $$PD/PC = \sin 25°.$$
>
> In the plane of P, D, and B, triangle PDB is a right triangle with a 65° angle at B (because PB is along an element of the cone). Therefore
>
> $$PD/PB = \sin 65°.$$
>
> The tangent principle gave us $PB = PF_1$, so we may substitute
>
> $$PD/PF_1 = \sin 65°.$$
>
> Consequently
>
> $$PF_1/PC = (PD/PC)/(PD/PF_1) = \sin 25°/\sin 65°.$$

We now have for the ellipse a characterization of a different character: The ellipse is the locus of points in a plane for which the ratio

(distance to a certain fixed point)/(distance to a certain fixed line)

is a certain constant *less than* 1. The line is called a **directrix** of the ellipse. [**Kline** (p. 96) writes that in the eight books that constitute his *Conics*, Apollonius never does talk in terms of foci or directrices.]

(iii) eccentricity



Now dance back to the figure (right) of the ellipse in its plane. Add to it the directrix $\mathcal{L}$ (solid black). We put it, as the figure in (ii) suggests, leftward of major vertex R. We made it vertical because it has to be parallel to the minor axis US: Since $UF_1 = SF_1$ and

UF_1/(distance U to $\mathcal{L}$) = SF_1/(distance S to $\mathcal{L}$),

we know U and S are equidistant from $\mathcal{L}$. Where exactly is $\mathcal{L}$?

(Symmetry tells us there must be a second directrix $\mathcal{M}$ to the right of T, and that the ratio

distance to F_2/distance to $\mathcal{M}$

must be the same constant. We need not analyze that one separately.)

> Let us evaluate that constant distance ratio. Write d for the distance from R to $\mathcal{L}$. Then R is $s - f$ from the focus F_1 and d from the directrix. For T, the distances are $s + f$ and $2s + d$. The two distance ratios have to match:
>
> $(s - f)/d = (s + f)/(2s + d)$.
>
> We can solve this equation for d. [In different words, *you* can solve this equation for d (Exercise 2).] Instead, let us use this fact: When two ratios are equal, the sum or difference of the numerators bears the same ratio to the sum or difference, respectively, of the denominators. For example,
>
> $3/5 = 18/30$
>
> allows us to conclude
>
> $3/5 = 18/30 = (3 - 18)/(5 - 30)$ (as well as $= [3 - 18]/[5 - 30] = [3 + 18]/[5 + 30]$).
>
> Using the differences, we conclude that the distance ratio is
>
> $(s + f)/(2s + d) \qquad = \qquad (s - f)/d \qquad = \qquad 2f/2s$.

[I understood that the sum of the numerators in a proportion bears the same ratio to the sum of the denominators when Jesse Douglas used the principle to give an elegant answer to a calculus problem. Prof. Douglas graduated from City College in 1916. He became world famous in the study of surfaces. In 1936, he and Lars Ahlfors received a special prize from the International Congress of Mathematicians in Oslo. That prize later acquired the name “Fields Medal,” despite the objections of its creator. (John Fields was Canadian, and intended the award to help heal the rifts in the international mathematical community resulting from World War I.) It became the highest international award in mathematics. If you saw *Good Will Hunting*, you heard Robin Williams describe it as akin to a Nobel Prize in math, except that it is granted only every four years to two, three, or four young mathematicians.]

The constant ratio of distance from focus to distance from directrix is that fraction f/s. The fraction is called the **eccentricity** of the ellipse. [We will use ε (epsilon) to denote it.] The name is appropriate, since f/s measures “off-centeredness.” It measures how far either focus is removed from the center, as compared with the unreachable maximum distance of s. What’s more, the eccentricity of an ellipse determines its shape. It is not just that small eccentricity means nearly circular shape, large eccentricity (close to 1) means elongation. Ellipses with equal eccentricity are *similar*: Their corresponding linear parts, like axes, focal lengths, chords joining vertices, or chords perpendicular to the major axis at the foci, are proportional. (See Exercise 3.)

Exercises III.A.7a

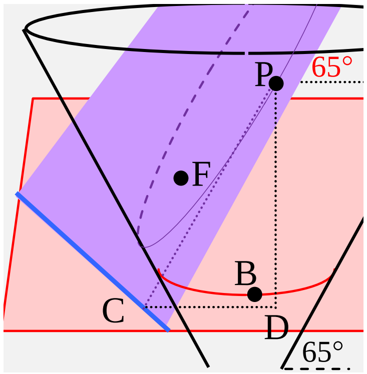
1. In the (last) face-on figure of the ellipse (iii), the width RT is intended to be 2 inches and the distance F_1F_2 between the foci 1 in. Exactly how much is the height US?
2. Solve
   $(s - f\,)/d \;=\; (s + f\,)/(2s + d)$
   for d in terms of s and f. (Hint: Douglas's idea makes it easier. Does the solution agree with the fact that R is $(s - f)$ from F_1 and d from $\mathcal{L}$?)
3. Suppose an ellipse has major axis 10 and minor axis 6.
   a) Show that its eccentricity ε is 4/5.
   b) Show that its **latus rectum** (the length of the chord perpendicular to the major axis through either focus) is
   $2(s - f\,)(1 + \varepsilon) \;=\; 18/5$. (A picture is essential.)
   c) Suppose a second ellipse has eccentricity 4/5. Show that its major axis $2M$, minor axis $2m$, and latus rectum L are in proportion to those of the given ellipse:
   $M/10 \;=\; m/6 \;=\; L/(18/5)$.
4. Characterize the **locus of Apollonius**, the set of points in a plane whose distance to one given point is a fixed multiple of its distance to a second given point. (Try using coordinate geometry in a specific example: Find an equation for the locus of points whose distance from the origin is 3 times their distance from (8, 0). Then put the equation into a form that allows you to give specific information about the locus.)
5. One of Kepler's laws says that the orbit of each planet is an ellipse, with the Sun at one focus. A planet is closest to the Sun when it reaches the major vertex closer to the Sun's focus, a point called "perihelion." It is furthest at the other major vertex, "aphelion."
   a) Earth's perihelion and aphelion distances are listed as 91.4 and 94.5 million miles. How long are the orbit's major axis, focal distance, and minor axis, and what is its eccentricity?
   b) The orbit of Mars is often described as "much more elliptical" than Earth's, because its eccentricity is around 0.093, vs. 0.017 for Earth. Show that its minor axis is about 99.6% as big as the major. [Accordingly, to human eyes, the orbit's non-circularity is not detectable.] (Hint: Try putting numbers into a sketch.)

c) the other sections

Recall from b(ii) that the focus-directrix description of the ellipse came from the relation

(distance to focus)/(distance to directrix) =
(sine of inclination of plane)/(sine of inclination of cone).

Nothing there demanded that the cutting plane be shallow. The relation holds equally true for the parallel section and the steep one.



(i) the parabola

For the parallel section, no sphere can sit on the cone above the cutting plane (violet in the figure at right). There is only the lower balloon (not shown), tangent to the cone along the red circle and to the cutting plane at a single focus F.

> Other than F, the points are labeled as before: P is a typical point on the section; B is where the element of the cone from P meets the circle of tangency; D is vertically below P, on the (pink) horizontal plane of

the circle; and C is the foot of the perpendicular from P to the (blue) line of intersection of the two planes.

By the tangent principle, PB = PF. From right triangles, we get

PD/PB = PD/PF = sin 65° and PD/PC = sin 65°.

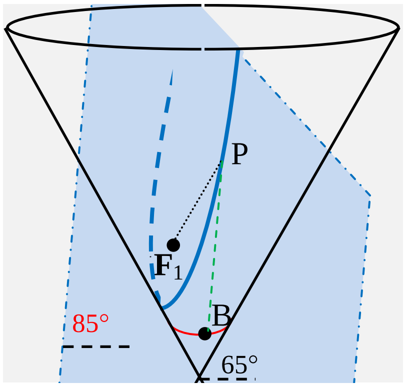
The parabola's distance ratio becomes

PF/PC = (PD/PC)/(PD/PF) = sin 65°/sin 65°.

That gives us two descriptions of the parabola. First in distance terms, the parabola is the locus of a point in the cutting plane whose distance to a certain fixed point (the **focus**) equals its distance to a certain fixed line (the **directrix**). Second, in eccentricity terms, we may as well refer to PF/PC as the **eccentricity** of the parabola. Therefore the parabola is that conic section whose eccentricity is 1.

(ii) the hyperbola

We can bring the same thinking to half the steep section.



View the cone's upper half in the figure at right. The cutting plane (blue) has inclination 85°. From point P on the upper branch of the hyperbola (heavy blue), the element PB (dashed green) of the cone reaches the circle of tangency (red) of the small sphere. Picture, as in the parabola figure above (but not shown here), the vertical PD to the plane of the red circle and the perpendicular PC to the line of intersection of the two planes. As before, right triangles yield

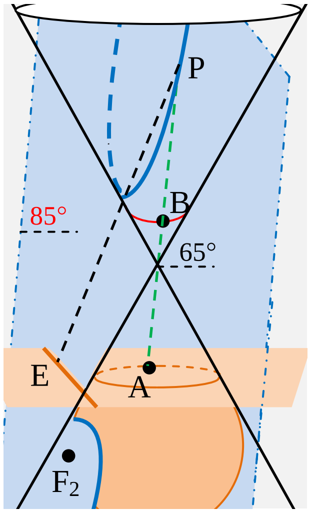
$PD/PB = PD/PF_1 = \sin 65°$ and $PD/PC = \sin 85°$.

The distance ratio for P becomes

$PF_1/PC = (PD/PC)/(PD/PF_1) = \sin 85°/\sin 65°$.

Thus, for every upper-branch point P, the distance to a certain fixed point in the upper half bears a (fixed) ratio bigger than 1 to the distance to a certain fixed line in the upper half.

Now extend that thinking to the lower half, even though the hyperbola's lower branch is not obviously a mirror image of the upper.



In the lower half of the cone, we can find one sphere (orange in this figure) tangent to the cone along the orange circle and to the cutting plane at F_2. Extend the element PB (dashed green) of the cone through the vertex to A on the orange circle. (If P and B are on the near side of the cone, then A has to be on the far side.) Let PE be the perpendicular to the line (heavy orange) of intersection of the orange circle's plane (light orange) and the cutting plane. Let PG (not shown) be the vertical to the orange plane from P. Again by the tangent principle, we have $PF_2 = PA$. The inclination of PA is the same as for PB, the inclination of the cone. Therefore

$PG/PA = PG/PF_2 = \sin 65°$.

The inclination of PE is that of the cutting plane. That forces

PG/PE = sin 85°.

It follows that

$PF_2/PE = \sin 85°/\sin 65°$.

For every upper-branch point P, the distance to a certain fixed point in the *lower* half bears the same fixed ratio to the distance to a fixed line in the lower half.

We conclude that the hyperbola has this focus-directrix characterization: There are *two* foci and two directrices, and for every point of the hyperbola, the ratio between distance to either focus and distance *to the corresponding directrix* ([near focus]/[near directrix] or [farther focus]/[farther directrix]) is a constant exceeding 1. That constant is the **eccentricity** of the hyperbola.

The last figure also gives us the distance characterization. Recall that

$$PF_1 = PB \qquad \text{and} \qquad PF_2 = PA.$$

Because B is between P and A,

$$PF_2 - PF_1 = PA - PB = BA.$$

That last is the constant distance *along the cone* from the level of the red circle to the level of the orange. Accordingly, the (complete, two-branch) hyperbola is the locus in a plane of a point whose distances to two certain fixed points differ (bigger distance minus smaller) by a certain fixed number.

(iii) the unification

The concept of eccentricity allows us a unified view of the conics as part of a continuum. At first, the parabola seemed to us to be a total break from the family of ellipses. Instead, in terms of eccentricity, the catastrophic change from ellipse to parabola is a smooth transition from eccentricity less than 1 to exactly 1. Similarly, the discontinuous change from parabola to hyperbola (for both of which, as for the ellipse, eccentricity determines shape) is just the continuation from eccentricity 1 to bigger.

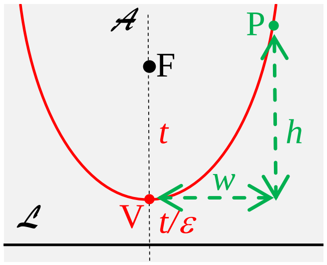
Moreover, Apollonius furthered this unification with a remarkable vision: 1800 years before Descartes, he used what amounts to Cartesian coordinates. He thought in terms of lengths (not strictly coordinates) to relate distances along the axis of a conic section to distances perpendicular to the axis.

Consider the conic with focus F, directrix $\mathcal{L}$, and points whose distances to F are eccentricity ε times their distances to $\mathcal{L}$. Build the figure at right in stages. The black parts are the focus, directrix, and the perpendicular $\mathcal{A}$ (for axis, dashed) from F to $\mathcal{L}$. Line $\mathcal{A}$ is necessarily an axis of symmetry for the conic. Call the focus-directrix distance *D*. Add now in red the point V (for *vertex*) located

$$t = D\varepsilon/(1+\varepsilon)$$

below F. For V, the distance to the directrix is

$$D - t \;=\; D(1-\varepsilon/[1+\varepsilon]) \;=\; t/\varepsilon. \qquad \text{(Verify that.)}$$



Hence V is on the conic, which must open upward to each side from V. Add last (green) the typical point P, located *h* above the horizontal line of V and *w* rightward from $\mathcal{A}$. Apollonius related *h* and *w*.

> Given that we plan to work with something like coordinates, let us make their use explicit. Place the origin of coordinates at V, with the *y*-axis up along $\mathcal{A}$. The focus gets coordinates (0, *t*), the directrix equation $y = -t/\varepsilon$, and the generic point P coordinates (*x*, *y*). The distance from P to F is
>
> $$PF = \sqrt{([x-0]^2 + [y-t]^2)}.$$
>
> From P to $\mathcal{L}$,it is $(y + t/\varepsilon)$. The focus-directrix characterization then reads
>
> $$\sqrt{([x-0]^2 + [y-f]^2)} \;=\; \varepsilon(y + t/\varepsilon).$$
>
> Square both sides and simplify to
>
> $$x^2 \;=\; 2(1+\varepsilon)ty + (\varepsilon^2 - 1)y^2. \qquad \text{(Likewise.)}$$

This last form is not as easy to read as our standard forms, but it captures in a single equation all four conic sections. “Four” includes the circle. That is remarkable, since the circle has eccentricity 0, for which the focus-directrix description and the algebra above are nonsense.

> If ε = 0, the form rearranges to
>
> $$x^2 + y^2 - 2ty \;=\; 0.$$
>
> Check that it describes a circle of radius *t* centered at F (Exercise 1a). If ε = 1, the rearrangement is
>
> $$x^2 \;=\; 4ty,$$

standard form for a parabola opening upward about the y-axis. For $0 < \varepsilon < 1$, we get

$$x^2 + (1 - \varepsilon^2)y^2 - 2(1 + \varepsilon)ty = 0.$$

You might recognize an ellipse with center up the *y*-axis (Exercise 1b). Last, $\varepsilon > 1$ yields

$$0 = (\varepsilon^2 - 1)y^2 + 2(1 + \varepsilon)ty - x^2.$$

This certainly represents a hyperbola, but you need to check that it conforms to our picture: specifically, that the transverse axis—the one that crosses the curve—is the *y*-axis (Exercise 1c)..

Finally, a historical note: That long form

$$x^2 = 2(1 + \varepsilon)ty + (\varepsilon^2 - 1)y^2$$

led Apollonius to create the names of the positive-eccentricity curves. For $\varepsilon < 1$, the equation looks like

$$x^2 = ry - sy^2, \qquad r \text{ and } s \text{ positive.}$$

The right side has something taken away, something lacking. The Greek word for “lack” or “deficiency” is *elleipsis*. [We write “…” to indicate that something has been left out; the symbol is called “ellipsis.”] Accordingly, Apollonius named the conic section of eccentricity less than 1—whose sectioning plane’s inclination is *short* of that of the side of the cone—an “ellipse.” For the contrary $\varepsilon > 1$, the form becomes

$$x^2 = ry + sy^2.$$

The right side has something added, an excess. The corresponding Greek word is *hyperbole* (or *huperbole*), for “excess” or “extravagance.” [Compare our word for “exaggeration.”] From it, we get the name for the section of eccentricity greater than 1, whose sectioning plane’s inclination *exceeds* the cone’s. Finally, if $\varepsilon = 1$, then we have

$$x^2 = ry.$$

The Greek *parabole* means “comparison” or “analogy.” [Our “parable” names a story that tells another story or gives a lesson by means of analogy. Compare **Kline**, pp. 92-93.]

Exercises III.A.7c

1. a) Transform

$$x^2 + y^2 - 2ty = 0$$

into standard form for a circle of radius *t* and center (0, *t*).

b) Show that for $0 < \varepsilon < 1$,

$$x^2 + (1 - \varepsilon^2)y^2 - 2(1 + \varepsilon)ty = 0$$

represents an ellipse with center at $(0, t/[1 - \varepsilon])$.

c) With $\varepsilon > 1$, put

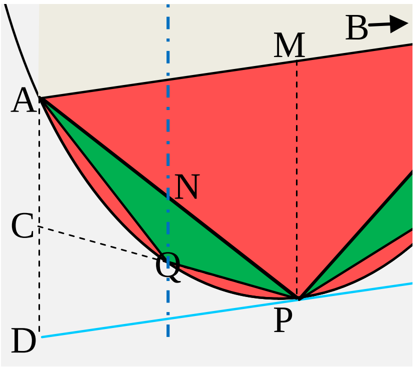
$$0 = (\varepsilon^2 - 1)y^2 + 2(1 + \varepsilon)ty - x^2$$

into a form for a hyperbola that crosses the *y*-axis.

2. We see that with properly placed and scaled axes, we can assign the equation $y = x^2$ to any parabola. Let A(a, a^2) and B(b, b^2) be the ends of a chord on a parabola so represented.

a) Suppose C(c, c^2) and D(d, d^2) are the ends of a chord parallel to AB. Show that the midpoints of both chords are on the same vertical line.

b) Return at right to the picture from the quadrature of the parabola (section III.A.6a(iv)). There, triangles APB and AQP are inscribed in the segments bounded by AB and AP, line PQ meets the vertical from A at C, and the tangent at P meets the vertical at D. Show that C is the midpoint of AD. (Hint: Show that the slope of PQ [you know their *x*-coordinates] is the average of the slopes of the tangent [same as that of AB] and PA. [Why does that suffice?])



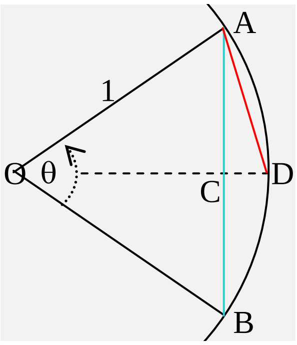
3. Show that in our generic conic, the latus rectum (the chord perpendicular to the axis through the focus) has length $L = 2(\varepsilon t + t)$. (Apollonius typically used L as the conic's parameter, rather than our $t = L/(2[\varepsilon + 1])$.)
4. a) Argue geometrically (informally) why, given a point on the side of the plane across the parabola from the focus—such as a point below a parabola that opens upward—there must be two tangents to the parabola from that point.
   b) (Calculus) Find the points of the parabola given by $y = x^2$ where the two tangents from the point (1, -3) meet the parabola.

8. Beyond the Golden Ages

a) trigonometry via chords

After Apollonius, Greek geometry entered a period of decline. In the years roughly 200 BCE to 600 CE, study concentrated on trigonometry. In the second century BCE, Hipparchus the astronomer (another Turk) developed an extensive body of trigonometric knowledge, including tables of trigonometric values, in pursuit of his occupation. More advances came later, including those of Claudius Ptolemy (Greek-Egyptian), also in the service of astronomy.

The Greek way in trigonometry was not like ours. They did not work with the right-triangle ratios—sine, cosine, and the others—that introduce our trig. They put values in terms of chords. In the figure at right, we see part of a unit circle, with a central angle AOB labeled θ. The length AB (blue) is the **chord** of the angle. (If the radius is not 1, then chord(θ) is AB/radius.)



You can relate the chord to our trig functions (Exercise 1). We are going to relate chords to other chords, to turn trigonometric values into arithmetic values. Those we could calculate, if we chose.

In the figure, we include the bisector (dashed) of angle AOB, meeting AB at its midpoint C and the circle at D. Angle AOD is therefore θ/2, and AD (red) is its chord. AOC is a right triangle with hypotenuse 1 and vertical leg AC = ½ chord(θ). That forces

$$\mathrm{OC} \quad = \quad \surd(\mathrm{OA}^2 - \mathrm{AC}^2) \quad = \quad \surd(1 - \tfrac14\,\mathrm{chord}^2(\theta))$$

for the horizontal leg. It leaves

$$\mathrm{CD} \quad = \quad \mathrm{OD} - \mathrm{OC} \quad = \quad 1 - \surd(1 - \tfrac14\,\mathrm{chord}^2(\theta)).$$

In right triangle ACD, therefore,

$$\mathrm{AD} \quad = \quad \surd(\mathrm{AC}^2 + \mathrm{CD}^2)$$

$$= \quad \sqrt{\Big(\tfrac14\,\mathrm{chord}^2(\theta) + 1 - 2\surd[1 - \tfrac14\,\mathrm{chord}^2(\theta)] + 1 - \tfrac14\,\mathrm{chord}^2(\theta)\Big)}.$$

We simplify to produce this "half-angle formula" :

$$\mathrm{chord}(\theta/2) \quad = \quad \sqrt{\big(2 - \surd[4 - \mathrm{chord}^2(\theta)]\big)}.$$

The figure intentionally has angle AOB ≈ 60°. Use it to see that chord(60°) = 1. From our formula,

$$\mathrm{chord}(30^\circ) \quad = \quad \sqrt{\big(2 - \surd 3\big)}.$$

Then $\quad \mathrm{chord}(15^\circ) \quad = \quad \sqrt{\big(2 - \surd[2 + \surd 3]\big)},$

$$\mathrm{chord}(7.5^\circ) \quad = \quad \sqrt{\big(2 - \surd[2 + \surd[2 + \surd 3]]\big)}$$ (Verify the last two.),

and the pattern is clear (Exercise 3).

Calculating those quantities requires time and adaptability.

Suppose we apply the square-root algorithm to $\sqrt{3}$ (Exercise II.B.3:5). We get the sequence:

estimate	partner	average
2	3/2	7/4
7/4	12/7	97/56
97/56	168/97	18,817/10,864.

It happens that 18,817/10,864 approximates $\sqrt{3}$ to within 3×10^{-9}. From the table alone, however, we can claim accuracy to only half the gap between 97/56 and 168/97, about 0.0001.

You can simplify the ongoing calculation by truncating:

$$18{,}817/10{,}864 \approx 1{,}881/1{,}086.$$

The truncated fraction approximates the longer one to accuracy better than 0.00001. Now write

$$2 - \sqrt{3} \approx 2 - 1{,}881/1{,}086 = 291/1{,}086 = 97/362.$$

(The reduction may not always be easy.) We approximate its square root as follows:

estimate	partner	average
1/2	194/362	750/1,448
750/1,448	140,456/271,500	407,005,288/786,264,000.

The long fraction approximates $\sqrt{(2-\sqrt{3})}$ to within 7×10^{-6}. The truncated 4070/7862 is within 0.00004 of the long fraction and is reducible. If you use 4070/7863, rounding the denominator upward after truncating, you stay equally close and achieve something desirable: an underestimate. The shortfall slightly offsets the overestimate that the square-root algorithm always produces. (Recall our elaboration of the need for underestimates in section III.A.6b(ii)).

You can see the combination of laborious computation and judicious adjustment needed to continue the process. Claudius extended it five more levels, to the value of chord(15/16°).

Carrying out these chord calculations had two purposes. One was to refine the approximation of π. Observe that [3 chord(60°)] is half the perimeter of the inscribed regular hexagon, as [48 chord(3.75°)] is the semiperimeter of the 96-gon. That was the target of Archimedes. You can see why with chord-oriented trigonometry, he used perimeters instead of areas. [See how he might have proceeded, using chords, in Appendix 1.] If you are willing to grind chord(15/16°) to eight decimal places, then you approximate the semiperimeter [192 chord(15/16°)] of the regular 384-gon to within 0.000002.

The other purpose was to serve the needs of trigonometry. We produced a "half-angle formula" for chords. You can similarly produce analogues for the double-angle formula (Exercise 4), sum formula, and other trigonometric identities. The sum formula (Exercise 5) would lead from chord(15/16°) to chord(30/16°), chord(45/16°), ... Thereby you would construct a table of trigonometric values. You could make the listing go by whole degrees, for example by beginning with the decent approximation

$$\text{chord}(1°) \approx 16/15\ \text{chord}(15/16°).$$

However, it is worth considering that the latter listing would not have been useful before the development of measuring instruments incorporating telescopes. Without optical aid, general angles would have been difficult to measure accurately. But certain angles can be *laid out* accurately. We already saw that the ancients could lay out precise right angles. You could build a 60°angle by stretching three equally long ropes into an equilateral triangle. You could accurately bisect an angle by laying equal ropes along its sides and marking the midpoint of the segment joining their ends (median in an isosceles triangle bisects the apex angle); or by stretching four equal ropes into a rhombus (the diagonal of a rhombus bisects the angles it joins). Bisecting repeatedly, you build good angles of 30°, 15°, 7.5°, and so on.

Exercises III.A.8a

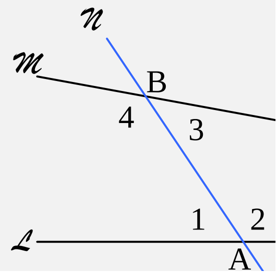
1. a) In this section's figure, use right triangle AOC to express AB = chord(θ) in terms of our trig ratios.
   b) Separately, use the law of cosines to relate chord(θ) to our trig ratios.
   c) Are the answers (a) and (b) equivalent?
2. Verify our half-angle formula for the cases:
   a) θ = 180°. You have to start by determining chord(180°) and chord(90°).
   b) θ = 120°. You need chord(120°) and chord(60°).
3. Show that
   $$\text{chord}(3.75^\circ) = \surd(2 - \surd[2 + \surd[2 + \surd\{2 + \surd 3\}]]).$$
4. Produce a "double-angle formula" by making $\alpha = \theta/2$ and solving
   $$\text{chord}(\theta/2) = \surd(2 - \surd[4 - \text{chord}^2(\theta)])$$
   for chord(2α) in terms of chord(α).
5. Produce a "sum formula" for chord($\alpha + \beta$) in terms of chord(α) and chord(β) . (Hint: Use the formula from Exercise 1(a).)

b) the parallel postulate

The salient purely geometric pursuit of the late Greek period was a chase that lasted about 2100 years. There was one particular postulate in Euclid's geometry that seemed to have character different from the others. It did not lead to contradictions or such evils. It just offended readers in a way that made them wonder whether it might be removed from the list of axioms in the manner we described in section III.A.5a, by turning it into a theorem. The two millennia spent in trying to prove it made the statement so famous that we will render it several different ways.

(i) Euclid's version

The postulate involves the lines $\mathcal{L}$ and $\mathcal{M}$, in one plane, cut by line $\mathcal{N}$ at two points A and B, as in the figure at right. We demand A ≠ B. That guarantees that we really have three different lines, and that if $\mathcal{L}$ and $\mathcal{M}$ intersect, then $\mathcal{N}$ does not share the common point. The **transversal** $\mathcal{N}$ (blue) forms eight angles at the two intersections. The four (numbered 1-4) between $\mathcal{L}$ and $\mathcal{M}$ are **interior** angles, the others **exterior** angles.



Euclid's Postulate. Suppose two lines are cut by a transversal so that interior angles on one side of the transversal add up to less than a straight angle. Then the two lines must meet on that side.

The figure is intended to make angle 3 < angle 1. Therefore

angle 2 + angle 3 < 180°.

According to the postulate, $\mathcal{L}$ and $\mathcal{M}$ must intersect somewhere to the right.

[I am not sure why geometers objected to this postulate. Maybe the reason is that it is not *local*. An earlier axiom, that you can draw a line joining two given points, is local. It looks at a specified region. Even if the points are light-years apart, you can visualize a long oval containing the two. By contrast, Euclid's Postulate makes a promise about the indefinite distance out to the right of the picture. In that way, it borders on talking about infinity.

It is wise to bear in mind that by "line," Euclid meant what we call "line segment." Watching over his shoulder for Zeno—as Eudoxus and Archimedes did—he did not visualize lines of infinite extent.

Instead, another earlier axiom said that “a line may be produced indefinitely.” That is, a segment may be extended by any finite number of copies of itself. “Then the two lines must meet …” is a promise about a sufficient but unstated number of extensions of segments.]

(ii) angles related to parallels

It is evident that when two lines are cut by a transversal, the four interior angles add up to two straight angles. Hence there are two possibilities. One is that the two on one side of the transversal sum to less than 180°. In that case, the two on the other side must sum to more than 180°, and Euclid says that the lines must meet on the less-than side. The other is that the two on each side add up to exactly 180°. We will describe the latter case in words that are more familiar. Then we will establish the inverse of Euclid’s postulate; that is, we will show that in this case, the lines *cannot* meet.

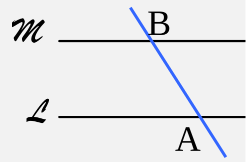
Suppose angle 2 + angle 3 = 180°. Then since angles 3 and 4 are supplementary,

angle 2 = 180° – angle 3 = angle 4.

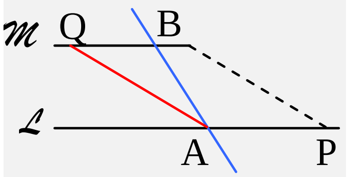
Two angles on opposite sides of the transversal are said to be **alternate**. Angles 2 and 4 are **alternate interior angles**. In the current case, both pairs of alternate interior angles must be congruent: From angle 2 = angle 4, we conclude that their supplements, angles 1 and 3, have to be congruent also.

Remember that lines in a plane are said to be **parallel** if they do not intersect.

Theorem 1. If a transversal to two lines forms congruent alternate interior angles, then the lines must be parallel.



For this theorem, you can make an intuitive appeal to symmetry. At right we have two lines cut by a transversal to form what look like 60° and 120° angles at both intersections. The picture is symmetric about the midpoint of AB. That is, if you rotate the page 180° about that point—for you, that means either rotating a monitor or rotating your head to view it—you end up with the same picture, a blue line sloping down to the right and crossing two others at 60° and 120° angles. Therefore if the two lines meet out toward the right, then they must also meet out toward the left. That would be a violation of one of the most basic properties we ascribe to straight lines, namely that two distinct lines cannot meet at two different points. We conclude that the lines cannot meet.



> For a more detailed proof, consider in the last figure the possibility that the extension of $\mathcal{M}$ toward the right (dashed in this figure) encounters $\mathcal{L}$ at point P. Reproduce length AP leftward from B, to point Q. Triangles BAP and ABQ are congruent by SAS. That is, they share side AB, side AP is congruent to side BQ, and the included angles BAP and ABQ are congruent alternate interior angles. Therefore angle BAQ matches angle ABP. That last angle is the interior angle on the right at B; remember that the line BP is simply $\mathcal{M}$. Therefore it matches the interior angle on the left at A. In other words, the angle between the left half of $\mathcal{L}$ and the transversal is the same as the angle between AQ and the transversal. Therefore $\mathcal{L}$ is the same line as AQ. That means $\mathcal{L}$ meets $\mathcal{M}$ at the point Q.
>
> From the assumption that the lines meet to one side, we have concluded that they intersect again on the other side. Therefore they cannot meet.

It is important to see that this theorem does not depend on Euclid’s Postulate. It follows from the idea of straightness embodied in the statement that two points determine a (single) line. That statement is a consequence of Euclid’s earlier postulates.

The theorem does guarantee that we can construct parallel lines: Since we can duplicate angles, we can *construct* congruent alternate interior angles, producing parallels.

(iii) equivalent postulates

Theorem 1 follows from the earlier postulates of Euclid. Its converse does not.

The Parallel Postulate. If two lines are parallel, then any transversal to them forms equal alternate interior angles.

The name "parallel postulate" is standard, but usually applied to our "Euclid's Postulate." The reason we have adopted it—for that matter, the reason we call it a postulate—is the next theorem.

Theorem 2. If you assume Euclid's postulate, then you can deduce the parallel postulate; and if you assume the parallel postulate, then you can deduce Euclid's postulate.

When each of a bunch of statements allows you to deduce all the others, the statements are said to be **equivalent**. The "equal value" they share is truth value: In every situation, they are all true, or else they are all false. We encountered equivalent statements in (ii), in the discussion just above Theorem 1. There, you can see that when lines are cut by a transversal, the statement

The interior angles on neither side of the transversal sum to less than a straight angle.

is equivalent to

The interior angles on each side sum to exactly a straight angle.

More generally, when a mathematical result takes the form

[This happens] if and only if [that happens].

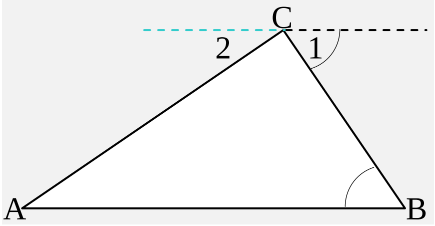
(as in Theorem 1 of section II.B.1), then the two bracketed statements are equivalent.

Proving the first part of Theorem 2 is easy.

> Look again at Euclid's postulate; we assume that statement to be true. We must prove that a transversal to two parallel lines forms equal alternate interior (hereafter **a/i**) angles. Suppose, then, we have a transversal to parallel lines. On either side of the transversal, the interior angles cannot add up to less than a straight angle, because then (by the assumed postulate) the lines would meet. Hence on each side, the interior angles must add up to a straight angle. By our observation just above, the a/i angles have to be equal. We have deduced the parallel postulate from Euclid's.

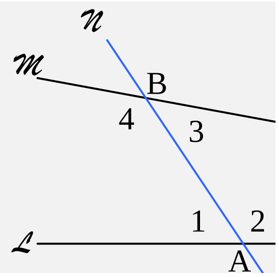
Proving the second part, converse to the first part, requires an intermediary that is itself of interest.

Theorem 3. (The Angle-Sum Theorem) If you assume the parallel postulate, then in every triangle, the angles sum to a straight angle.



> To the right, we see triangle ABC. We reproduce angle B at C, using BC as one side and the dashed black half line as the other. That gives us angle B = angle 1. The dashed green line extends the dashed black to the left of C. This new line and line AB are cut by transversal BC to form equal a/i angles. By Theorem 1, the added line is parallel to AB. If we assume the parallel postulate, then we infer that the transversal AC must also form equal a/i angles. Thus, angle A matches angle 2. Then the sum of the angles of the triangle is
>
> angle A + angle ACB + angle B = angle 2 + angle ACB + angle 1.
>
> It is obvious that the latter three add up to a straight angle.

Now we deduce Euclid's postulate from the parallel postulate.



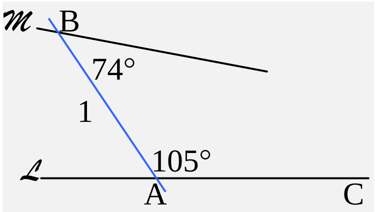
> Assume the parallel postulate; that is, assume that whenever two lines are parallel, every transversal to them forms equal a/i angles. Suppose that in the figure at right, angles 2 and 3 sum to 175°. Then
>
> angle 4 = 180° – angle 3
> = 180° – (175° – angle 2) = (angle 2) + 5°.
>
> The a/i angles do not match. By the assumption, the lines $\mathcal{L}$ and $\mathcal{M}$ cannot be parallel; they meet someplace. That place is not to the left. If they met to the left, then $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{N}$ would enclose a triangle whose angles would be angle 1, angle 4, and one other. But angles 1 and 4 already add up to
>
> angle 1 + angle 4 = 360° – (angle 2 + angle 3) = 185°.
>
> The (assumed) parallel postulate guarantees that no triangle can have angle sum that high. Therefore the meeting place is on the side where the interior angles have the smaller sum. That proves Euclid's postulate, on the assumption of the parallel postulate.

Taking stock, we see that Euclid's postulate and the parallel postulate amount to the same statement. We see also that either of them implies the angle-sum theorem. It happens that the angle-sum theorem implies the others, but that is harder to establish. Let us pretend we have established the implication. With it, we establish the theorem as a third statement equivalent to the two postulates. At the same time, it is important for us to remember that we did not *prove* Euclid's postulate, or the parallel postulate, or the angle-sum theorem. What we showed is that if you want to prove one using the other postulates of Euclidean geometry, then it suffices to prove either of the others.

That brings us back to this section's introduction. The first well-known attempt to write a proof of Euclid's postulate from the others was by Proclus (Alexandrian, 410-485 CE. Most of what we [believe we] know about Thales and Pythagoras is from the testimony of Proclus.) The proof turned out to depend on the principle that parallel lines are necessarily equidistant. When we (teaching) introduce parallel lines, we sometimes use the analogy of train tracks. That analogy is based on this principle. You can prove the equidistance, *if you assume the parallel postulate* (Exercise 2). Indeed, like the angle-sum theorem, the equidistance turns out to be equivalent to the postulate. In other words, what Proclus hoped was a proof had the fatal flaw that *it assumed what it was trying to prove*. That malady, called **circular reasoning**, afflicted all the other attempts to prove the postulate in the 1400 years following Proclus. See Exercise 3 for a sample of it.

Exercises III.A.8b

1. How did Greek geometry change after the time of Apollonius (roughly 200 BCE)?
2. Prove that parallel lines are equidistant: Draw two parallel lines and put points A and B on one of them; prove (assuming the parallel postulate) that the perpendiculars from A and B to the other line are equally long.
3. a) Draw triangle PQR with side PQ = 1, angle P = 74°, and angle Q = 105°. Find the length of QR.

   b) At right is a partial reproduction of this section's opening figure, with the transversal making angles of 74° and 105° to the lines $\mathcal{M}$ and $\mathcal{L}$, respectively, and the distance AB set at 1. Point C is chosen so that

   AC = sin 74°/sin 1°.

   Prove that triangle BAC is congruent to triangle PQR in (a).

   c) Prove that $\mathcal{M}$ coincides with line BC. This proves that $\mathcal{M}$ must meet $\mathcal{L}$ at C.

d) You could use the argument in (a)-(c) as a template to build a proof of Euclid's postulate. Show that the "proof" would be circular, by spotting the places where the postulate is a hidden assumption. (There are at least three such places in the argument. Recall "hidden assumptions" from section III.A.5b.)

Section III.B. Number Theory

The Greeks did not make any great contributions to numeration, but they did advance the study of numbers. They had two names for that study. One was *logistiki*, for what we would call "arithmetic." They seem to have held a low opinion of the crass study of computation. What they held in high esteem was *arithmoi*, the more dignified study of properties of numbers that we would call "number theory." Here we do a great deal of the latter. Just about all of it appeared in *The Elements*. Euclid's great work was not limited to geometry.

1. The Fundamental Theorem

We start with a principle that must have been known to plenty of ancient people. Its proof may have been given first by the Pythagoreans.

The Fundamental Theorem of Arithmetic. Every natural number beyond 1 can be factored into the product of primes, and the **prime factorization** is unique.

To elaborate the Theorem, we need the elementary notion of divisibility. Let a and b be integers. They can be positive or negative, but it helps to demand $a \neq 0$. We say a **divides** b if there is an integer k with $b = ka$. Thus, 5 divides 65 because $65 = 13 \times 5$. Evidently 13 also divides 65. If a divides b, then we also say a **is a factor** of b, a **is a divisor** of b, b **is divisible by** a, b **is a multiple** of a.

Remember that a natural number is called **prime** if its only positive divisors are 1 and the number itself. For convenience, we do not count 1 as a prime number. You can check that 5 and 13 are both prime, and we have seen that 65 is not. We count each of 5 and 13 as a "product of primes" having a single factor. Also, we say "unique" to abbreviate "unique except for the order of the prime factors." Accordingly, $65 = 5 \times 13 = 13 \times 5$ is "the unique factorization" of 65.

a) prime factorization

You cannot prove general statements by resort to examples. However, we will adopt the practice of using specific numbers to reduce the abstractness of arguments—really, to keep the number of variables down—to the extent that we can still clearly illustrate how to structure a general proof. In that spirit, we give evidence for the part of the Theorem preceding the comma, using the example of 360.

> We try dividing 360 by the numbers from 2 to 359. Actually, we need only try to divide by primes (Exercise 2), and we have to try the primes only from 2 to $\sqrt{360}$ (Exercise 3). If none of those primes divides 360, then we conclude that 360 is prime and is therefore its own factorization.
>
> We immediately find that 2 divides 360. We obtain
>
> $$360 = 2 \times 180.$$
>
> That expresses 360 as the product of a prime and another factor. The other factor can be no greater than 360/2, because we are dividing by at least 2. Repeating on the end factors, we proceed through
>
> $$\begin{aligned} 360 &= 2 \times 2 \times 90 && \text{(two primes and a factor no greater than 360/4)} \\ &= 2 \times 2 \times 2 \times 45 && \text{(three primes and a factor no greater than 360/8)} \\ &= 2 \times 2 \times 2 \times 3 \times 15 && \text{(four primes and a factor no greater than 360/16)} \\ &= 2 \times 2 \times 2 \times 3 \times 3 \times 5 && \text{(five primes and a factor no greater than 360/32).} \end{aligned}$$
>
> It happens that we have arrived at a prime factorization. Notice, though, that the process could go on

> for an unpassable maximum of three more steps. If there were three more steps, the end factor would be at most 360/256; it would be 1. The process of breaking down into primes necessarily terminates.

In practical terms, factorization usually works better if we make the factors closer together than, say, 2 and 180. If for example we choose the fairly obvious

$$360 = 36 \times 10,$$

we can then write

$$\begin{aligned} 360 &= 6 \times 6 \times 5 \times 2 \\ &= 2 \times 3 \times 2 \times 3 \times 5 \times 2. \end{aligned}$$

We end up as before with a factorization consisting of three 2’s, two 3’s, and a lone 5.

Exercises III.B.1a

1. Assume 36 divides the integer b. Show that:
   a) If $b > 0$, then $36 \leq b$. (Best attack: induction. We will cover that method in 1900 years.)
   b) 36 divides every multiple of b.
   c) If 36 also divides c, then 36 divides both $b + c$ and $b - c$.
2. Given an integer, show that its smallest divisor beyond 1 must be prime.
3. The number 12,079 is not prime. Show that the smallest prime that divides it is no more than $\sqrt{12{,}079} \approx 110$.

---

b) integer division

Proving the uniqueness part of the Fundamental Theorem requires some results that are of independent interest to us.

Theorem 1. (The Division Algorithm) Suppose d (as in **divisor**) is a natural number and n is an integer. Then there exist two integers q (as in **quotient**) and r (as in **remainder**) such that $n = qd + r$, and these integers are unique if we require that

$$0 \leq r < d.$$

Notice the restriction on r. It is allowed to be zero, but must be *strictly smaller* than the divisor; see Exercise 1. Notice also that if (and only if) $r = 0$, then d divides n.

> In support of Theorem 1, pick $d = 18$ and $n = 1000$. Some multiples of 18, like 1000 × 18, exceed 1000. By a property of the natural numbers, there must be a *smallest* such multiple. Call that multiple $18i$. (What *is* the actual value of i?) Then $18(i - 1)$ does not exceed 1000. Put all that as
>
> $$18(i - 1) \leq 1000 < 18i.$$
>
> (That line is really what underlies the Division Algorithm: Every integer is between two consecutive multiples of 18; because the gap between those multiples is 18, the integer has to be from 0 to 17 above the lower multiple.) From the previous inequality, we have
>
> $$0 \leq [1000 - 18(i - 1)] < 18i - 18(i - 1) = 18.$$
>
> Accordingly, the equation
>
> $$1000 = 18(i - 1) + [1000 - 18(i - 1)]$$
>
> tells us that $(i - 1)$ is a quotient and $[1000 - 18(i - 1)]$ a remainder upon division of 1000 by 18.

[We referred to the following property of the natural numbers: Every nonempty set of natural numbers has a smallest member. The property is called the **well-ordering principle**. We need to accept it for now, on the promise that we will justify it eventually.]

As for uniqueness, suppose 1000 equals both $18q + r$ and $18Q + R$. From

$$18q + r = 18Q + R,$$

we have

$$18(q - Q) = R - r.$$

That says 18 divides $R - r$. If R and r are between 0 and 17, then $R - r$ is between –17 and 17. In that string of integers, the only multiple of 18 is 0. The reason is that for 18, the positive multiples are 18 and higher, the negative multiples -18 and lower (Exercise 1a above). Therefore $R - r$ is zero:

$$R = r.$$

That forces $18q = 18Q$, and therefore $q = Q$. The quotient and remainder are unique.

It is odd that Theorem 1 makes a statement with the name "algorithm." An **algorithm** is normally defined as a step-by-step method for doing some job. The steps must be carried out in sequence, although an instruction may say "jump now to step x." The algorithm must leave no decisions open; it must anticipate any that can arise and specify how to make them. (In 1996, failure of an algorithm—a vast computer program—to specify how to react to a calculation, turned an Ariane rocket into an expensive, unguided flying pile of ordnance.) By anybody's definition, the algorithm must guarantee that it will reach an answer. The answer does not have to be good news; it could be, "What you want to do is impossible." But termination must be guaranteed. Additionally, the definition sometimes provides that the algorithm must specify, or imply, the maximum time or number of steps it will need to terminate. We have seen such specifications. Look back at our factorization of 360 (subsection (a)). There we observed that it takes no more than 8 cycles of finding new prime factors, because $360/2^8$ is already less than 2; and each cycle requires no more than $\sqrt{360}$ divisions to produce the new factor.

However, we forgive the odd name. The Division Algorithm is a remarkable combination of elementariness and power. It is clearly elementary; we introduce it in the schools by fourth grade. You will see its power in our frequent uses of it. It underlies a number of important principles. One of them comes right now, on the way to proving the uniqueness part of the Fundamental Theorem.

Theorem 2. If a prime divides a product, then it has to divide one of the factors.

It is not true that if a *number* divides a product, then it divides some factor. For example, 6 divides 9 × 10, but it does not divide either of 9 or 10. Theorem 2 describes a property of *primes*.

To prove Theorem 2, we need another property of primes.

Theorem 3. If p does not divide n, then there are two integers i and j such that $pi + nj = 1$.

Fix two integers a and b. The expression $ai + bj$, i and j understood to be integers, is called an **integer combination** of a and b. (You could as well call it an integer combination of a and j, or i and b, or i and j.) The combinations of a and b include a and b themselves,

$$a = a(1) + b(0) \qquad \text{and} \qquad b = a(0) + b(1),$$

their sum and difference

$$a + b = a(1) + b(1) \qquad \text{and} \qquad a - b = a(1) + b(-1),$$

and $\quad 0 = a(0) + b(0).$

Theorem 3 says that if p does not divide n, then some integer combination of p and n has a value of 1.

To illustrate proof of Theorem 3, observe that the prime 19 does not divide 1000. The integer combinations of 19 and 1000 include some positive ones. (Name one.) By the well-ordering principle, one of those must be smaller than all the others. Let

$$d = 19i + 1000j$$

be the smallest positive combination of 19 and 1000.

This number d divides both 19 and 1000. To see why, apply the Division Algorithm to 19 with d:

$$19 = qd + r.$$

> Rewrite that as
>
> $$r = 19 - qd$$
> $$= 19 - q(19i + 1000j) = 19(1 - qi) + 1000(-qj).$$
>
> You see that r is an integer combination of 19 and 1000. Since r has to be less than d and d is the smallest positive combination, r cannot be positive. It has to be 0. That is, d divides 19. By exactly the same reasoning, d divides 1000.
>
> The statement that d divides 19 narrows the possibilities. Only 1 and 19 divide 19; d has to be either 1 or 19. But it is not 19, because d divides 1000 and 19 does not. Therefore d is 1. We have in d an integer combination of 19 and 1000 whose value is 1. (Find one in Exercise 3.)

We will continue to talk in terms of *integer* combinations, allowing ourselves the indulgence of using negative coefficients. If $19i + 1000j = 1$, then either i or j has to be negative. The Greeks did not have a conception of negative numbers. What we write with a negative j, they would have rendered as the positive difference of positive multiples

$$19i - 1000(-j) = 1.$$

If instead it is i that is negative, then it would have been

$$1000j - 19(-i) = 1.$$

Our combinations provide a symmetry that lets us avoid having to deal with $19k - 1000l$ and $1000m - 19n$ as separate cases.

> Keep in mind that if we can produce one form of *difference*, then we can produce the other. For the reasonable numbers $p = 5$ and $n = 18$, we easily see that
>
> $$18(2) - 5(7) = 1.$$
>
> Think of that as saying that some multiple of 18 is 1 more than some multiple of 5. Multiply by $5 - 1 = 4$ to get
>
> $$18(8) - 5(28) = 5 - 1,$$
>
> then rearrange to
>
> $$1 = 5 + 5(28) - 18(8) = 5(29) - 18(8).$$ (Get the calculator and verify.)
>
> Now some multiple of 5 is 1 more than some multiple of 18.
>
> Had we started with that last equation, we could have multiplied by $18 - 1 = 17$ to write
>
> $$18 - 1 = 5(29 \times 17) - 18(8 \times 17),$$
>
> which rearranges to
>
> $$18(1 + 8 \times 17) - 5(29 \times 17) = 1.$$ (Calculator?)

Now we justify Theorem 2.

> Assume 19 divides the product $1000m$ and does not divide 1000. Then some combination $19i + 1000j$ equals 1. Multiply
>
> $$1 = 19i + 1000j$$
>
> by m to write
>
> $$m = m(19i + 1000j) = 19(mi) + 1000m(j).$$
>
> On the right, 19 obviously divides $19(mi)$. It also divides $1000m(j)$, because the latter is a multiple of $1000m$ and 19 divides $1000m$ (previous Exercise 1b). Therefore 19 divides the sum
>
> $$19(mi) + 1000m(j) = m$$ (previous Exercise 1c).
>
> We have shown that if 19 divides the product $1000m$ and does not divide the first factor, then it has to divide the other. That is our evidence for Theorem 2.

The statement extends to any number of factors. If 19 divides the product $abcd$, then it divides a, or else has to divide bcd. In the latter case, it has to divide b, or else cd. If it comes down to that last, then it has to divide either c or d. If a prime divides a product, then it must divide at least one of the factors.

c) uniqueness of factorization

Finally we can establish that 360 has just one factorization.

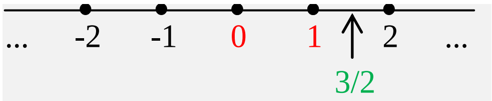
> Suppose 360 also factors as *pqrs*.... From
>
> $$2 \times 2 \times 2 \times 3 \times 3 \times 5 \quad = \quad pqrs...,$$
>
> we see that 2 divides the product *pqrs*.... Therefore 2 must divide one of the primes *p*, *q*, *r*, *s*, ... That means 2 has to *be* one of those primes. Make it *p*: 2 = *p*. Then we may cancel to write
>
> $$2 \times 2 \times 3 \times 3 \times 5 \quad = \quad qrs....$$
>
> Again we conclude that one of those primes *q*, *r*, *s*, ... is 2, so that (say)
>
> $$2 \times 3 \times 3 \times 5 \quad = \quad rs....$$
>
> You see how this continues. We keep canceling one prime on the left with one on the right. Neither side can run out of primes before the other, because then one side would equal 1 while the other held a product of primes. Therefore the list *p*, *q*, *r*, *s*, ... consists of exactly six primes, comprising three 2's, two 3's, and one 5. Prime factorization is unique.

Exercises III.B.1c

1. What are the quotient and remainder on division by 5 for:
   a) 38 b) -38?
2. Assume 1000 divides integers *b* and *c*. Prove that 1000 divides every integer combination of *b* and *c*.
3. Find an integer combination of 19 and 1000 that equals 1. (One attack: The division algorithm will give you a combination equal to 12. Double that one, you get 24; subtract 19, you get 5; quadruple, you get 20; subtract 19.)
4. Write $360 = 2^3\, 3^2\, 5^1$. This expression is called the **prime-power factorization** for 360.
   a) Assume *d* divides 360. Show that the prime-power factorization of *d* must be
   $d = 2^i\, 3^j\, 5^k$,
   with no other primes and with the powers satisfying $0 \le i \le 3$, $0 \le j \le 2$, $0 \le k \le 1$.
   b) May the powers be 0, 0, 0, respectively? May they be 3, 2, 1?
   c) In view of (a) and (b), how many divisors does 360 have, counting 1 and 360?
   b) In general, how many divisors are there for the number with factorization $p^a\, q^b ... r^c$?

2. Irrational Numbers

The Pythagoreans made a discovery that threw Greek geometry into something of a crisis. It is helpful to put it into a modern context.



Recall that in our teaching of the real number line [better said, in my teaching of the number line], we begin by assigning two real numbers to two points on the line. It does not matter which two numbers we use; as soon as we assign them, we specify orientation on the line (which direction the numbers increase) and a unit of measure. For convenience, then, we choose to start by assigning 0 and 1 (red in the figure). Lay off their distance to the right, and it makes sense to attach to the point so reached the next natural number. At the same distance further right, we assign the number after that, and so on. The line also extends to the left, so we assign integers downward from 0 to the uniformly placed points leftward.

Evidently this process leaves in-between points. Fortunately, we have in-between numbers. For example, to the midpoint of the segment joining the points labeled 1 and 2, we assign the number halfway between 1 and 2, shown in green. In each gap between integer-labeled points, we assign the infinity of rational numbers between the corresponding integers to the infinity of points. The question naturally arises: Does that take care of all the points on the line?

For the Greeks (who would not have put in 0 and the part of the line left of there) the answer would have been affirmative. We know that you can take any integer length *m* and "multisect" it into *n* equal parts, thereby constructing the fractional length *m*/*n*. If you take literally the description of a length less than 1 as a "fraction," then you are picturing those lengths—all of them—as coming from breaking a whole number length into another whole number of parts. That seems to be how the Greeks thought of *all* the intermediate lengths, until the Pythagoreans showed that the answer is no.

Remember that it is possible to construct a length whose square is 2. We can do it by constructing either an isosceles right triangle with unit legs, on whose hypotenuse the square is 2; or a square whose area is 2 units (Exercise III.A.3:2a), whose side therefore fits the bill. In our notation, we denote the length by $\sqrt{2}$. Here is what the Pythagoreans discovered:

Theorem 1. There is no fraction whose square is 2.

For proof, we take a small departure from the usual argument.

> If there were a fraction *m*/*n* with
>
> $$2 = (m/n)^2 = m^2/n^2,$$
>
> then we would have
>
> $$m^2 = 2n^2.$$
>
> That equation cannot hold between natural numbers. Look at it in terms of prime factorization. Whatever the factorization of *m* is, the factorization of m^2 necessarily consists of the same primes written twice as many times. (For example,
>
> $$360 = 2\times2\times2\times3\times3\times5 \qquad \text{forces} \qquad 360^2 = 2\times2\times2\times3\times3\times5 \times 2\times2\times2\times3\times3\times5.)$$
>
> Therefore the factorization of m^2 has an even number of primes. The same of course goes for any square, including n^2. Therefore the two integers m^2 and $2n^2$ cannot be equal; the latter factors into an odd number of primes, owing to the extra 2. (Why can't a product of an odd number of primes somehow contrive to match the product of an even number of primes?) The equation
>
> $$2 = (m/n)^2$$
>
> cannot be satisfied by any fraction *m*/*n*.

We take it for granted that the real numbers encompass the fractions and the unfractions (hereafter **rational** and **irrational numbers**.) But it was their knowledge of the *natural* numbers that led the Pythagoreans to discover that there *are* reals, like $\sqrt{2}$, besides the rationals.

Exercises III.B.2

1. a) Prove that if a natural number is not a square (of another natural number), then its square root is irrational.
b) Prove the same for any root: If *n* is not the *k*'th power of another natural number, then the *k*'th root of *n* is irrational.
c) More generally, prove that the lowest-terms rational number *m*/*n* has a rational *k*'th root iff each of *m* and *n* is itself the *k*'th power of an integer. (Examples: 8/27 has a rational cube root, because $8 = 2^3$ and $27 = 3^3$; but 8/12,345 does not have a rational fourth root, because 8 is not a fourth power. Are you sure 8/12,345 is in lowest terms?)

3. Eudoxus and the Nature of Ratio

The discovery of irrational numbers was one reason philosophers, Zeno being perhaps best known, began to demand more precise definitions and tighter logic. Greek geometers found themselves in need of a close look at the very nature of numbers and ratios of numbers. Ratios, especially equality of ratios, are at the heart of the relations we get from similarity. Those include, for one big example, the statement that the circumferences of circles are in proportion to their radii. The geometers needed someone who could address both the geometric tradition and the numerical Oriental tradition. They needed Eudoxus.

Eudoxus gave the definition that two (positive) quantities **have a ratio** if some multiple of each exceeds the other. Under this criterion, there is a ratio between $\sqrt{2}$ and 5:

5×1 exceeds $\sqrt{2}$, because $(5 \times 1)^2$ is more than $(\sqrt{2})^2$, and
$\sqrt{2} \times 4$ exceeds 5, because $[\sqrt{2} \times 4]^2 = 2 \times 16$ is more than 5^2.

He then defined *equality* of ratios as follows: Two **ratios are equal** if (equal) multiples of the two numerators relate the same way to (perhaps other, equal) multiples of the denominators.

The last is a rarity, a definition that is easier to understand in symbols than in words.

> In symbols:
> $a/b = c/d$
> means that whenever the multiple ma is less than the multiple nb, then $mc < nd$; and whenever $ma = nb$, then $mc = nd$; and whenever $ma > nb$, then $mc > nd$.

Notice that both definitions use comparisons in terms of integer multiples. Dividing in the equality definition by n, we can cast it as:

> The ratios a/b and c/d are equal iff every rational multiple $(m/n)a$ is less than, or equal to, or greater than b, alike as $(m/n)c$ is below, at, or above d.

Either way, the comparison is in terms of the good old numbers, meaning natural numbers and their ratios. The equality definition has the desirable quality of actually looking at the relative sizes of a and c compared to b and d. View that against our familiar criterion,

$a/b = c/d$ iff $ad = bc$.

This test puts the equality of the ratios in terms of the equality of two products that are blind to the ratios as quantities and even to the relation between a and b or between c and d.

The other definition ("quantities have a ratio") is worth some study, because it leads to a statement that helps confirm our picture of the real numbers. The statement bears the name of Archimedes, but Archimedes himself traced it back to Eudoxus.

The Axiom of Archimedes. If r and a are two positive real numbers, then some multiple of a exceeds r.

You can think of it in words as saying that if you take enough equal steps, no matter how small, rightward along the real line, then you eventually pass any fixed point on the line, no matter how far. In the language of Eudoxus, it says that any two positive real numbers have a ratio.

To relate it to our picture of the reals, set $a = 1$. The axiom says that if r is positive, then some multiple of 1—in other words, some natural number—is to the right of r. If s is a negative real number, then $-s$ is positive, there exists some natural n to the right of $-s$, and $-n$ is to the left of s. Therefore the integers are coextensive with the reals; there is no real number that lies beyond the rightward reach or beyond the leftward reach of the integers. In more detail, since there are natural numbers to the right of r, the well-ordering principle says that there is a *least* such natural number. Call it m. Then $m - 1$ is not to the right of r. That is,

$m - 1 \leq r < m$.

Every real r, then, is itself an integer or lies in the gap between consecutive integers. We may study the vicinity of r by studying instead the interval between 0 and 1.

The vicinity of 0 is crowded with fractions. We can prove that if ε is any positive real, no matter how close to 0—no matter how small—there exist unit fractions between ε and 0 (Exercise 2). [We used ε to represent eccentricity; now we make the standard use of it, to represent a positive number presumed to be small.] On the basis of that, we can establish a remarkable property of the rational numbers.

Theorem 1. The rational numbers are **dense** in the real line: If *r* and *s* are any two real numbers, then it is possible to find a rational number between them.

Let us proceed assuming *r* and *s* are both positive and $r < s$. If they are of opposite signs, then 0 is between them; and if they are both negative, the argument is simply the mirror image of what we write.

> Look at the length $s - r$ between them. By Exercise 2, there is a unit fraction $1/n$ with
>
> $$0 < 1/n < s - r.$$
>
> By the axiom of Archimedes, some multiple of $1/n$ exceeds *r*. By the well-ordering principle, there is a smallest such multiple; denote it by $m(1/n) = m/n$. That means $(m - 1)/n$ does not exceed *r*. Of necessity,
>
> $$(m - 1)/n \le r < m/n.$$
>
> How far above *r* is m/n? Put our two inequalities together and you have
>
> $$\begin{aligned} m/n &= (m - 1)/n + 1/n \\ &\le r + 1/n < r + (s - r) = s. \end{aligned}$$
>
> The rational number m/n is between *r* and *s*.

We see that our “real line” model of the real numbers reflects ideas that go back to Eudoxus.

Exercises III.B.3

1. Use Eudoxus’s definition to show that
   $\sqrt{2/5} = \sqrt{8/10}$.
2. Let ε be a positive real number. Prove that there is a unit fraction $1/n$ between 0 and ε. (Hint: $1/\varepsilon$ is a positive real number.)
3. Prove that between any two real numbers, there exists *an infinity* of rational numbers.

4. Euclid and Number Theory

Here we capture some of the number theory covered in Books 7-10 of the *Elements*.

a) common divisors

Much of this section applies to any finite set of integers. However, we will limit our talk to *pairs* of integers.

A natural number that divides each of *a* and *b* is called a **common divisor** of them. Recall that there are always common divisors; it is automatic (in math: **trivial**) that 1 is a common divisor of any pair of integers. Sometimes that is it. If 1 is the only common divisor of *a* and *b*, then we say *a* and *b* are **relatively prime**. Since a divisor cannot exceed (the absolute value of) what it divides (Exercise III.B.1a:1a), there must be among the common divisors a biggest one. We call that one the **greatest common divisor** (hereafter **GCD**) of *a* and *b*. Thus, the GCD of 35 and 75 is 5, because the only divisors of 35 = 5 × 7 are 1, 5, 7, and 5 × 7, and of those only 1 and 5 divide 75. For a like reason, 35 and 76 are relatively prime.

[Remember that we choose to say of relatively prime numbers that they have “no common divisor.”]

Here are two key results.

Theorem 1. The smallest positive integer combination of two integers is their GCD.

For evidence, we work with 640 and 1000. Let

$$d = 640i + 1000j$$

be the smallest possible combination you can make from them. (Remember why it is that *there must be* a smallest positive combination.) We leave as Exercise 4 the proof that d must divide 640 and 1000. It then has to be the biggest of the common divisors: If k is another common divisor, then k divides all the combinations of 640 and 1000 (Exercise III.B.1c:2), including d. From the fact that k divides d, we conclude $k \leq d$. Therefore d is the biggest common divisor.

Theorem 2. (The Euclidean Algorithm) Let a and b be natural numbers. Divide a into b to get a remainder r_1; divide r_1 into a to get remainder r_2; and so on, until you get a zero remainder. The last legal divisor—meaning either a or the last nonzero remainder—is the GCD of a and b.

To continue with our example, divide 1000 into 640 to get

$$640 = 0 \times 1000 + 640.$$

Next we divide 640 into 1000, and absorb the valuable lesson that you save time by dividing the smaller into the larger in the first place. We proceed through

$$\begin{aligned} 1000 &= 1 \times 640 + 360 \\ 640 &= 1 \times 360 + 280 \\ 360 &= 1 \times 280 + 80 \\ 280 &= 3 \times 80 + 40 \\ 80 &= 2 \times 40. \end{aligned}$$

We have reached a zero remainder. That was inevitable; the algorithm *must* reach 0 remainder, because the remainders decrease. The Euclidean algorithm says that 40 is the GCD of 640 and 1000.

To establish the algorithm, observe that the last line says 40 divides 80. The previous line says

$$280 = 3 \times 80 + 40.$$

Because 40 divides 80, it also divides the integer combination $3 \times 80 + 40$. That is, 40 divides the divisor and dividend on that previous line, 80 and 280. The line two up from last reads

$$360 = 1 \times 280 + 80.$$

We know that 40 divides 80 and 280. Therefore 40 divides the divisor and dividend on that line, 280 and 360. Keep climbing the ladder, and see that no matter how high it goes, we must arrive at 40 dividing the original divisor and dividend, 640 and 1000. The Euclidean algorithm always produces *some* common divisor.

Is it the biggest? Unwind the algorithm up from the next-to-last line. Rewrite the line as

$$40 = 280 - 80(3).$$

Substitute 80 from the line before,

$$80 = 360 - 280(1),$$

to get

$$40 = 280 - 3[360 - 280(1)] = 360(-3) + 280(4). \qquad \text{(Check it!)}$$

Substitute 280 from the next line up, and …. Read the pattern: This process keeps showing 40 as an integer combination of the guests of honor on each line, going up line by line. It continues

$$\begin{aligned} 40 &= 360(-3) + [640 - 360(1)](4) = 640(4) + 360(-7) \\ &= 640(4) + [1000 - 640(1)](-7) = 1000(-7) + 640(11). \qquad \text{(Check!)} \end{aligned}$$

The algorithm shows that 40 is an integer combination of 640 and 1000. (More than that: It delivers the coefficients that express 40 as such.)

On one hand, 40 is a common divisor, so that $40 \leq$ GCD. On the other hand, the GCD is the *smallest* of the positive integer combinations, implying GCD ≤ 40. We conclude that 40 is the GCD of 640 and 1000, and that the Euclidean algorithm always produces the GCD.

Remember that our use of negative numbers is consistent with our way but not the Greek. In terms of positive numbers, we have

$$40 = 640(11) - 1000(7).$$

To express 40 as a multiple of 1000 minus a multiple of 640, multiply through by $1000/40 - 1 = 24$:

$$1000 - 40 = (1000/40 - 1)40 = 640(11 \times 24) - 1000(7 \times 24),$$

then rearrange to get

$$1000(7 \times 24 + 1) - 640(11 \times 24) = 40.$$

Exercises III.B.4a

1. Use the Euclidean algorithm to find the GCD of 360 and 94, and express the GCD as an integer combination of the two. Check with a calculator.
2. Show that 999,999,999 and 1,000,000,001 are relatively prime. Then find an integer combination that equals 1.
3. a) Describe how to obtain the GCD of two natural numbers, given their prime factorizations.
   b) Factor 360 and 94 into primes and apply (a) to obtain their GCD.
   c) Describe how to get the least common multiple (exactly what it sounds like, abbreviated **LCM**) of two numbers from their factorizations.
   d) Find the LCM of 360 and 94.
   e) Show that for any two numbers, GCD times LCM is the product of the numbers.
4. Let $d = 640i + 1000j$ be the smallest positive integer combination of 640 and 1000. Mimic the proof of Theorem 3 in III.B.1b to show that d divides both numbers.
5. Show that if k divides the product ab and is relatively prime to a, then k must divide b. (One approach: Mimic the proof of Theorem 2 in III.B.1b. This exercise is a generalization of that theorem. Contrast the statement with our remark that 6 divides 9 × 10, even though it divides neither 9 nor 10. On the other hand, 9 × 10 = 5 × 18, and 6 is relatively prime to 5.)

b) Pythagorean triples

We now have the tools we need to keep the promise to establish the structure of Pythagorean triples, as given in Theorem 1 of section II.B.1. The first half says that if u and v are relatively prime and of opposite parity, then they yield a primitive triple.

> Let $u > v$ be any natural numbers. It is automatic that the three numbers
>
> $$u^2 - v^2,\ 2uv,\ \text{and}\ u^2 + v^2$$
>
> form a Pythagorean triple (Exercise 1).
>
> Suppose that the triple is not primitive; that is, the three numbers have a common divisor. Then they must share a common *prime* divisor p. Necessarily, p also divides the sum
>
> $$(u^2 + v^2) + (u^2 - v^2) = 2u^2$$
>
> and the difference
>
> $$(u^2 + v^2) - (u^2 - v^2) = 2v^2.$$
>
> One way that could happen is if $p = 2$. In this case, $u^2 + v^2$ and $u^2 - v^2$ are both even, and u and v must have the same parity. The only other way p can divide both $2u^2$ and $2v^2$ is for p to divide both u^2 and v^2, and therefore divide both u and v. (Reasons?) Then u and v are not relatively prime.
>
> Summarizing via the contrapositive: If u and v are relatively prime and of opposite parity, then the triple has to be primitive.

[Before we give evidence that every primitive triple has this form, you should consider how anybody could have thought up such a complicated structure. A good bet: The process must have been what we used to facilitate construction of the hexagon (section III.A.4b) and the pentagon (III.A.2c); namely, you imagine what you want to study, then deduce properties that allow you to produce it.]

Start, then, by letting a, b, c form a primitive Pythagorean triple: They have no common divisor and

$$a^2 + b^2 = c^2.$$

We saw (Exercise II.B.1:1c) that a and b cannot both be odd. They also cannot both be even, because if they were, then all three numbers would be. In fact, no two of a, b, and c can have a common divisor.

If say b and c had a common divisor, then they would share a common prime divisor q. In that case, q would also have to divide

$$a^2 = c^2 - b^2,$$

would therefore divide a. Similar reasoning works if anything divides both a and c or both a and b.

For illustration, take a primitive triple like 93, 476, 485. (Check that they form a triple.) Observe that the smaller numbers are one odd and one even; and

$$93 = 3(31), \qquad 476 = 2^2(7)17, \qquad \text{and} \qquad 485 = 5(97)$$

are pairwise relatively prime.

(When among three or more numbers, no two have common divisors, the numbers are **pairwise relatively prime**. It is not automatic that relatively prime numbers are pairwise prime. The three numbers 15, 21, 35 do not have a common divisor, but 15 and 21 do, 21 and 35 do, and 15 and 35 do.)

To establish that 93, 476, 485 is given by the Babylonian method, write

$$\begin{aligned}(476/2)^2 &= 476^2/4\\ &= (485^2 - 93^2)/4\\ &= (485 - 93)/2\ \ (485 + 93)/2.\end{aligned}$$

The first fraction and the last two are whole numbers, because their numerators are even. Therefore the product of the two numbers $(485 - 93)/2$ and $(485 + 93)/2$ is a square. Those two numbers have no common divisor: If they did, then that divisor would also divide their sum and difference

$$(485 - 93)/2 + (485 + 93)/2 = 485,$$
$$(485 + 93)/2 - (485 - 93)/2 = 93,$$

two numbers we know to be relatively prime.

Because $(485 - 93)/2$ and $(485 + 93)/2$ are relatively prime and their product is a square, they both have to be squares. A product of non-squares can be a square, like 2×8, but not if the factors are relatively prime. For proof, assume mn is a square with m and n having no common divisor. First, each prime in the factorization of mn appears an even number of times. Second, since m and n are relatively prime, the primes in mn come from m or from n, *but not both*. That means the primes in m appear in m an even number of times, likewise n. Hence m and n are squares.

Having shown the two fractions to be squares, we denote

$(485 + 93)/2$ by u^2,
$(485 - 93)/2$ by v^2.

(Verify that those fractions really are squares.) We already saw that u^2 and v^2 are relatively prime, and therefore so are u and v. Furthermore,

$$485 = u^2 + v^2,$$
$$93 = u^2 - v^2,$$
$$(476/2)^2 = u^2\,v^2.$$

The first two equations guarantee that u and v cannot have the same parity. The last one gives

$$476/2 = uv,$$

and $476 = 2uv$. The Babylonian formulation gives all possible primitive triples.

Exercises III.B.4b

1. Show that for any natural numbers $u > v$, the three numbers
   $u^2 - v^2$, $2uv$, and $u^2 + v^2$
   constitute a Pythagorean triple.
2. a) For the 6, 8, 10 triple, find u and v such that $8 = u^2 - v^2$, $6 = 2uv$, and $10 = u^2 + v^2$.
   b) Show that the same cannot be done for 9, 12, 15. Is that a contradiction?

c) commensurability

Euclid, being of the Greek tradition, always associated numbers with lengths. The arithmetic notion of divisor goes with the geometric notion of measure. Length *d* **measures** length *a* if *a* is a whole number of times *d*. If *d* measures both *a* and *b*, it is a **common measure**. The biggest such thing is the **greatest common measure**.

> Any two rational lengths a/b and c/d have a common measure $1/(bd)$. (That statement is true for integers too: set $b = d = 1$.) The reason our criterion for equality of ratios,
>
> $a/b = c/d$ iff $ad = bc$,
>
> works is that it compares the fractions as multiples of the common measure $1/(bd)$:
>
> $a/b = [ad](1/bd)$ and $c/d = [bc](1/bd)$.
>
> The Euclidean algorithm applies to rational numbers exactly as to integers. For 2/5 and 8/3, we have
>
> $8/3 = 6 \times 2/5 + 4/15$.
>
> (The quotient 6 comes from $(8/3)/(2/5) = 40/6 = 6+$, and the remainder from $8/3 - 6 \times 2/5 = 4/15$.) Next,
>
> $2/5 = 1 \times 4/15 + 2/15$,
>
> $4/15 = 2 \times 2/15$, no remainder.
>
> By the algorithm, 2/15 is the greatest common measure.

One irrational can measure another. Thus, $\sqrt{2}$ measures $\sqrt{50} = 5\sqrt{2}$. Lengths or numbers with common measures are said to be **commensurable**. But no irrational is commensurable with any rational.

> If $\sqrt{2}$ and 5 had a common measure *d*, so that say
>
> $\sqrt{2} = md$ and $5 = nd$,
>
> then we would have $\sqrt{2}$ as the fraction
>
> $\sqrt{2} = m(5/n) = (5m)/n$.

When numbers are commensurable, they have a "ratio," as in "rational." From

$8/3 = 20 \times (2/15)$ and $2/5 = 3 \times (2/15)$,

we judge that the two fractions are in the ratio 20/3. We see why Zeno, who preceded Hippocrates, would have objected to talk that $\sqrt{2} - 1$ and $1 - \sqrt{2}/2$ are in the "ratio" $\sqrt{2}/1$ (which we said in squaring the lune in section III.A.3b). We see also why Eudoxus had to develop the new notion of ratio to compare incommensurables like $\sqrt{2}$ and 5.

d) primes and perfects

There are two theorems credited to Euclid himself. One says that the set of primes is infinite. Greek philosophers avoided or even opposed the idea that a collection could be abundant beyond any number, perhaps because the capacity of the concept of infinity to generate paradoxes is—how shall we say—infinite. But Euclid gave a proof that the primes are without end.

Theorem 1. Suppose $p_1, p_2, p_3, \ldots, p_k$ is a list of primes. Then some prime number is not on the list.

The proof is remarkably brief.

> Let p_1, p_2, p_3, ..., p_k be a list of prime numbers. Add 1 to their product: Write
> $$N = (p_1\, p_2\, p_3 \ldots p_k\,) + 1.$$
> If this number is prime, then it is a prime not on the list, because N exceeds all of p_1, p_2, p_3, ..., p_k. If instead it is not prime, then it is divisible by some prime number. This prime is not on the list, because none of p_1, p_2, p_3, ..., p_k divides N: The expression for N shows that on division by any of those primes, N has remainder 1. Either way, we verify that some prime has been left off the list.

It is worth comparing what Theorem 1 says to what we said Euclid proved. The theorem indicates that if you make a list of 10^{1000} primes, then it is possible to find one that is not listed. You could add that one to your list, but then you would have a list with $10^{1000} + 1$ primes. The theorem would equally apply to this enhanced list, would therefore guarantee that the enhanced list still does not have all the primes. In other words, no finite collection of primes exhausts the supply. It is in that sense that Theorem 1 establishes that the collection of primes is infinite.

Our way of stating the theorem puts it into the important mathematical class of **existence theorems**. The statement says that under some circumstance, something with a specified property exists. It does not state what that something is. Rather, it cites circumstantial evidence that there is a smoking gun somewhere with our something’s fingerprints on it. Contrast that with the statements:

1. On a given line segment, there exists a point that breaks the segment into the golden section.
2. There exists a length whose square is 2.

We did not just state that the point and length exist; we showed how to construct them. Indeed, providing evidence of how you produce the something in question is called **constructive proof**.

The other theorem ascribed to Euclid deals with perfect numbers. A natural number is called **perfect** if it is the sum of its **proper** divisors, the divisors smaller than the number itself. Thus, 6 and 28 are perfect: 6 is divisible by 1, 2, and 3, and 1 + 2 + 3 = 6; 28 is divisible by 1, 2, 4, 7, and 14, and equals their sum. An equivalent definition is that the sum of *all* the divisors is twice the number.

Theorem 2. Suppose $2^n - 1$ is prime. Then $2^{n-1}(2^n - 1)$ is perfect.

We work as usual with examples.

> $2^1 - 1 = 1$ does not count.
> $2^2 - 1 = 3$ is prime, so $2^1(2^2 - 1) = 6$ is perfect.
> $2^3 - 1 = 7$ is prime, so $2^2(2^3 - 1) = 28$ is perfect.
> $2^5 - 1 = 31$ is prime, so $2^4(2^5 - 1) = 496$ is perfect.
>
> The skipped one,
> $$2^4 - 1 \;=\; (2^2 - 1)(2^2 + 1),$$
> is not prime. The same goes for every even n; $2^n - 1$ factors as the difference of squares. For that matter, n cannot even be **composite** (the opposite of prime): $2^{15} - 1$ factors as both the difference of cubes
> $$(2^5)^3 - 1 = [(2^5) - 1][(2^5)^2 + (2^5) + 1]$$
> and the difference of fifth powers
> $$(2^3)^5 - 1 = [(2^3) - 1][(2^3)^4 + (2^3)^3 + (2^3)^2 + (2^3) + 1].$$
> [Check both those multiplications. I do not always tell the truth. For example, the sentence before this one is false.]

Let us use the example $n = 5$ to give evidence for the general statement in Theorem 2.

Since $2^5 - 1$ is prime, $2^4(2^5 - 1)$ is its own prime power factorization. (See Exercise III.B.1c:4.) Hence any divisor of $2^4(2^5 - 1)$ has to be the product of between zero and four 2's and between zero and one $(2^5 - 1)$'s. Accordingly, we can list the divisors:

$$\begin{array}{lllll} 1 & 2 & 2^2 & 2^3 & 2^4 \\ 1(2^5-1) & 2(2^5-1) & 2^2(2^5-1) & 2^3(2^5-1) & 2^4(2^5-1). \end{array}$$

Add them up and notice that we can factor out $[1 + 2 + 2^2 + 2^3 + 2^4]$. We produce the sum

$$s \quad = \quad [1 + 2 + 2^2 + 2^3 + 2^4][1 + (2^5 - 1)].$$

Now we use one of the interesting properties of the powers of 2: They add up to one less than the next power. In symbols,

$$1 + 2 + 2^2 + ... + 2^n = 2^{n+1} - 1 \qquad \text{(Exercise 3).}$$

Therefore the divisors of $2^4(2^5 - 1)$ add up to

$$\begin{aligned} s &= [2^5 - 1][1 + (2^5 - 1)] \\ &= [2^5 - 1]\,2^5 \quad = \quad 2 \times 2^4(2^5 - 1). \end{aligned}$$

The divisors add up to twice the number; $2^4(2^5 - 1)$ is perfect.

Remarkably, 2300 years after Euclid, it is still unknown whether this formulation accounts for *all* of the perfect numbers. It is known that it takes care of all the *even* perfects, but no one has settled the question of whether there are odd ones. And for these Euclidean perfects, unlike the situation with the primes, no one knows whether there is an infinity of them.

Exercises III.B.4d

1. a) Use the Euclidean algorithm to find the greatest common measure of two line segments of lengths 2/7 and 20/3.
   b) Answer the same question by multiplying both lengths by any common denominator, finding the GCD of the resulting integers, then dividing by the denominator.
2. (**Boyer**) The number $2^{13} - 1$ is prime. Use this fact to find the related perfect number.
3. Prove that
   $$1 + 2 + 2^2 + ... + 2^n = 2^{n+1} - 1.$$
   (One approach: $x^{n+1} - 1$ has a standard algebraic factorization.)

5. Algebra

The biggest Greek contribution to algebra came, like the investigations of Proclus into the parallel postulate, centuries after the golden ages. The *Arithmetica* of Diophantus (around 250 CE) introduced two new features: It took an abstract, exact approach, and did not tie it to geometry. Diophantus created a whole class of algebraic problems. He brought original insights to their treatments, including the ability to deal with multiple specifications (what for us are simultaneous equations).

Diophantine equations come under the heading nowadays called "indeterminate equations." An indeterminate equation carries insufficient information to pinpoint an answer. Rather, it is satisfied by some set of numbers, sometimes an infinite set. "Solving" it means describing the set of solutions. This kind of problem is not completely new to us:

$$x^2 + y^2 = z^2, \qquad x, y, \text{ and } z \text{ required to be integers,}$$

is a Diophantine equation. We solved it by characterizing the Pythagorean triples. We mention three more examples:

$$x^3 + y^3 = z^3, \qquad\qquad x^3 + y^3 = z^3 + w^3,$$

and the (simultaneous) system

$$x = 10y + 4 \qquad x = 21z + 5,$$

all to be satisfied by integers. We will have reason to come back to them later.

To treat one particular kind of example, look at the form

$$ax + by = c$$

("the general linear equation"). Here a, b, and c are given, fixed integers, x and y are integers to be determined. If a and b have a common divisor, then that factor has to divide c. Thus,

$$60x - 250y = 73$$

has no solutions (Exercise 1). If instead the equation reads

$$60x - 250y = 730,$$

then we divide by the GCD of 60 and 250 to produce

$$6x - 25y = 73.$$

This is an **equivalent** equation (same solutions) in which the coefficients are relatively prime. Think of it as asking for a multiple of 6 that is 73 more than some multiple of 25, and you hark back to something we considered before.

> With 6 and 25 relatively prime, some integer combination equals 1. In different words, the equation
>
> $$6i - 25j = 1$$
>
> has a solution. We know we can solve that one via the Euclidean algorithm. However, with such small numbers, think multiples. To find a multiple of 6 that is 1 more than a multiple of 25, look at
>
> $$25 + 1, \qquad 50 + 1, \qquad 75 + 1, \qquad \ldots.$$
>
> The first of those divisible by 6 is 126. Thus,
>
> $$6(21) - 25(5) = 1.$$
>
> Now multiply through by 73 to write
>
> $$6(21 \times 73) - 25(5 \times 73) = 73.$$
>
> One solution of $6x - 25y = 73$ is
>
> $$x = 21 \times 73 = 1533, \quad y = 5 \times 73 = 365.$$ (Verify, then do Exercise 2.)

Exercises III.B.5

1. Show that there are no integers x and y with
   $$60x - 250y = 73.$$
2. a) Find one solution of
   $$6x - 25y = 858.$$
   b) Find one solution of
   $$21x - 15y = 300.$$

Section III.C. The Astronomers

In astronomy, unlike mathematics, the Greeks were heirs to the Babylonians.

Much of man's knowledge in astronomy owed to the movements of the Moon. [I will frequently call it "Luna."] The ancients could see that the stars are fixed to the dome of heaven. Even over the course of *twenty* human lifetimes, human eyes could not detect any relative motion among the stars, motion say that would make a change in the shape of Orion or the Scorpion. (It was not until about 300 years ago that observatories, equipped with telescopes and fine measuring instruments, were able to confirm minute relative movements.) But there were things that did move relative to the stars. The Greeks called them *astere planetai*, heavenly objects that *wander*. They included Luna, the Sun, and five starlike objects that of course we now call "planets."

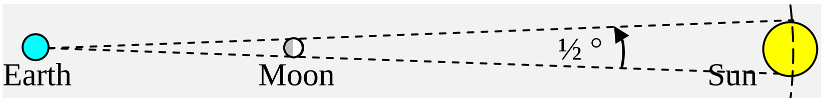
Of those, Luna is clearly closest, because it "occults" (passes in front of) the others. A little more on average than twice a year, it occults the Sun. (Those events are called "solar eclipses," which is something of a misnomer. "Eclipse" really means shadowing, not blocking.) A few times a year, it occults a visible star. Every few years, it occults one of the planets. Moreover, it goes around the sky (from the viewpoint of Earth) faster than the others, in a time that accounts for the word “month.” After that, the fastest swing around the sky is the Sun’s yearly trip. The ancients took that as evidence that Sun is next closest. The planets other than Venus and Mercury—the planets that are not always near the Sun—require years: Mars takes 2+, Jupiter 12-, Saturn 29+. Those, then, might be above the Sun. With the stars exhibiting no motion, and since the planets also occult stars sometimes, it was reasonable to conclude that the stars lie on a higher level than the planets.

One more lunar event was important. About as frequently as solar eclipses—often half a moon (two weeks) before or after a solar eclipse—Luna is eclipsed. The fact that those events happen when Luna and the Sun are opposite in the sky suggests that a lunar eclipse is Luna’s entry into Earth’s shadow. (Those eclipses, plus the near-blackness of the Moon at solar eclipses, say that Luna shines not by its own light, but by reflected sunlight.) The fact that Earth’s shadow upon the Moon is always circular, no matter what lunar eclipse you watch, suggests that Earth is spherical. Even in antiquity, then, those with astronomical knowledge knew that Earth is spherical. They could have further figured that Luna is also a sphere, illuminated on one side, so that its phases are due to how much of the lighted half we can see.

1. Aristarchus

Aristarchus, like Pythagoras born on Samos, lived something like 310-230 BCE, so that he overlapped decades with Archimedes. He is best known for proposing a Sun-centered universe. But he was a wonderful geometer, and here we will see the ingenuity he brought to the attempt to determine distances in the heavens.

a) size compared with distance



Long before the Greeks, people knew that the Sun and Moon show us faces of equal angular size. The figure illustrates the equality, for which clear evidence came from solar eclipses. At those eclipses, Luna is on average just big enough to cover the Sun. As for the angle, according to **Boyer**, Archimedes ascribed the approximate ½ ° determination to Aristarchus. That seems odd, because rope stretchers should have long before been able to determine the angle using marks on the ground or on structures. In any case, we will calculate using the ½ ° value.

[Luna’s size varies considerably, owing to its changing distance to Earth. We now know that Luna’s distance covers a range of almost 14%, closest point to farthest. (That’s “perigee” to “apogee”. Compare “perihelion” and “aphelion” in Exercise III.A.7:5.) The Earth-to-Sun distance varies as well, but more tamely: 4% perihelion to aphelion.]

The equal angular size implies a quantitative relationship between the sizes and distances to Sun and Moon. Say Luna has radius *r* and distance *l* from Earth, Sun has radius *R* and distance *L*. From the figure, we see that for each of them, the ratio diameter/distance is

$$2r/l \;=\; 2R/L \;=\; \text{chord}(1/2\,°).$$

Aristarchus could have approximated the chord as (8/15 chord[15/16 °]). (See section III.A.8a.) More likely, he would simply have used the arc of a half-degree angle (dotted arc within the Sun in the figure). That arc (on a unit circle) would be

$$1/2\;(2\pi/360) \;\approx\; (22/7)/360 \;=\; 11/1{,}260.$$

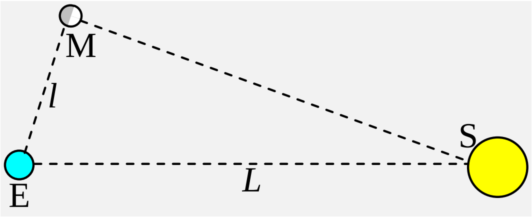
Therefore we have

$$2r/l \;=\; 2R/L \;=\; 11/1{,}260.$$

We may also say that each of Sun and Moon is 1,260/11 ≈ 115 times as far as it is big.

> When you deal with small angles, there is little difference between the angle's sine, its chord, and its radian measure, which by definition is the length of the arc the angle intercepts on the unit circle. For a 60° angle, we know the sine is $\sqrt{3}/2 \approx 0.866$, the chord is 1, and the arc is $2\pi/6 \approx 1.047$. That is a spread of about 21% from smallest to biggest. When you get down to 0.5°, the sine is 0.00872654 [via scientific calculator], the chord is 0.00872662 [via $\text{chord}(\theta) = 2\sin(\theta/2)$, which answers Exercise III.A.8a:1a], and the arc is $2\pi/720 \approx 0.00872665$. That spread is less than 0.0013%.

b) the ratio of the distances



Now advance the clock a quarter of a moon, seven days, to turn the previous figure into this one. (The calendar calls this situation "First Quarter.") From Earth (E), we see half the sunlit hemisphere of Luna (M). Aristarchus realized that at this time, the angle SME between the Sun-Moon line and Moon-Earth line is 90°. Therefore the distance l from Earth to Moon is related to the distance L from Earth to Sun by

$$l/L = \cos(\text{angle SEM}).$$

Aristarchus tried to measure angle SEM and obtained 87°. [Remember that we are in modern idiom; Aristarchus would have used neither cosine nor degrees.] That puts the ratio of distances at

$$l/L \;=\; \cos(87°) \;=\; \sin(3°)$$
$$\approx\; \text{arc of } 3° \;=\; 3(2\pi/360) \;\approx\; 1/19.$$

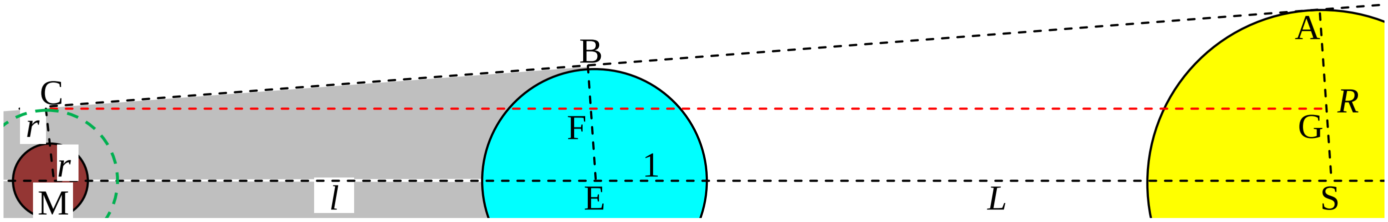
Since $2r/l = 2R/L$ implies

$$r/R = l/L,$$

it follows also that Luna has 1/19 the size of the Sun.

c) size and distance in terms of Earth

Finally, we let seven more days pass and arrive at the figure below. There, we have the full Moon (maroon) in eclipse, in the gray cone of shadow cast leftward by Earth (blue). The centers M of the Moon and E of Earth are l apart. The center S of the Sun (yellow) is in the line ME, at distance L from E. (That alignment is a statistical impossibility, as Exercise 1c suggests, but allows the illustration Aristarchus made.) Luna has radius r, Earth has radius 1, and the Sun has radius R. The common tangent to Earth and Sun is AB, whose extension leftward is the edge of the shadow. The ancients estimated that it took Luna about the same time to enter Earth's shadow completely as it took, after entering, to leave completely. The timing suggested that Earth's shadow cone, at the distance where Luna enters it, is about twice the size of Luna. Accordingly we have drawn the outline (dashed green) of a sphere concentric with Luna and having radius $2r$. The extension of AB is tangent to the outline at C.



Notice also that we have drawn AB sloping upward from left to right; in other words, we made the Sun larger than Earth. If Sun were smaller than or equal to Earth in size, then AB would slope down to the right or be horizontal, and the shadow would be greater than or equal to the Sun. That would make the shadow 19 or more times the size of the Moon. It would imply that at every full Moon, Luna would undergo a multi-day total eclipse. The calculations of Aristarchus dictated that Sun is larger, indeed significantly larger, than Earth. [Maybe the idea that a small Earth orbiting a large Sun makes more sense than Sun orbiting Earth was part of the reason Aristarchus proposed a heliocentric universe.]

The figure has one more important element. We draw the (red) horizontal at C, meeting BE at F and AS at G. We *made* CFG parallel to MES, but it is important to note that AS, BE, and CM are also parallel. They are radii, and therefore perpendicular to the common tangent ABC. Consequently, we (believers in the parallel postulate) conclude that triangles CFB and CGA are similar, and that CMEF and CMSG are parallelograms.

> The similarity gives us
> $\quad$ BF/FC = AG/GC.
> From the CMEF parallelogram, we get
> $\quad$ BF = BE – FE = 1 – 2r and FC = l.
> From CMSG, we have
> $\quad$ AG = R – 2r and GC = l + L.
> The similarity proportion becomes
> $\quad$ $(1 - 2r)/l = (R - 2r)/(l + L)$.
> Putting that together with the three other relationships
> $\quad$ $2r/l = 2R/L = 11/1{,}260$ and $l/L = 1/19$,
> we find (Exercise 2)
> $\quad$ $r = 20/57 \approx .35$, $l \approx 80$, $R \approx 6.7$, $L \approx 1{,}530$.

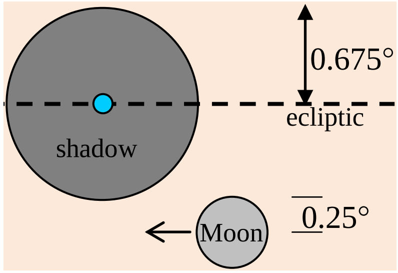
Thus, Aristarchus calculated that Luna is about one-third the size of Earth and 80 Earth radii (“40 Earths”) distant, Sun the size of seven Earths and 770 Earths away.

d) modern values

Our current data say that the angular size of the Sun is on average about 0.53°. That changes the diameter to distance ratio to $(0.53 \times 2\pi/360) \approx 1/108$. Exercise 2 shows that this change, surprisingly, has no effect on the calculated r, though of course it reduces l. The data also say that the size of Earth’s shadow is closer to 2.7 times the size of Luna. Using that value, instead of 2, reduces the calculated r to about .28 the size of Earth. More fundamental is that we now know the ratio between the distances is actually close to 400. That change reduces the r estimate by little, to about .27. However, it implies a Sun as big as 108 Earths and about 11,700 Earths away.

The 400 ratio means that the cosine of angle SEM is 1/400, which makes the angle about 89.86°. Aristarchus underestimated how short angle SEM is of a right angle by a factor of more than 20. The determination error is understandable for two reasons. First, it is guesswork to find the moment when the Moon is precisely half illuminated. Second, in trying at that moment to measure the angle, you need to aim something at the center of a high, blinding Sun. You cannot wait for Sun to be near the horizon, because then the Sun you see is actually an optical illusion caused by atmospheric refraction; and obviously you are helpless when it is below the horizon. Still, keep in mind the brilliance and *mathematical* validity of what **Boyer** calls the “unimpeachable” method Aristarchus used.

Exercises III.C.1



1. This problem puts the frequency of lunar eclipses in terms of probability. In the figure, the dashed line is the ecliptic. Earth's dark gray shadow is a circle with a radius of 0.675°, centered on the line. The light gray Moon has radius 0.25° (1/2.7 the shadow's size) and moves horizontally to the left. It moves so that its center follows a random horizontal level within a band that reaches 5° above and 5° below the ecliptic.
   a) Show that the probability of Luna's catching some part of the shadow is 1.85/10. Under this model, an eclipse, maybe only partial, happens on average every 10/1.85 ≈ 5.4 moons.
   b) Show that the probability of Luna's crossing entirely within the shadow is 0.85/10. According to this, we should average a total eclipse every 10/0.85 ≈ 12 moons.
   c) Show that the probability of Luna's center crossing the central 10% of the shadow (the blue target in the middle of the shadow, with radius 0.0675°) is 0.135/10 ≈ 1/74. Accordingly, an eclipse that dark comes about once in 74 moons ≈ six years.
2. Use the size and distance relations
   $2r/l = 11/1260, \qquad 2R/L = 11/1260, \qquad r/R = 1/19$
   to solve
   $(1 - 2r)/l \;=\; (R - 2r)/(l + L)$
   for r. (Notice that in the solution process, the diameter-to-distance ratio 11/1260 cancels; r is independent of it.) Then find the corresponding l, R, and L.

2. Eratosthenes

You can see that putting absolute numbers to the relative calculations of Aristarchus required determination of the size of Earth. That was provided by Eratosthenes (275-194, one or two generations after Aristarchus).

In mathematics, the name of Eratosthenes is generally known for "the sieve," a method for compiling lists of primes. Imagine a list of integers from 2 to someplace,

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 ....

We underline 2, then cross out every second number after it, to produce

$\underline{2}$ 3 ~~4~~ 5 ~~6~~ 7 ~~8~~ 9 ~~10~~ 11 ~~12~~ 13 ~~14~~ 15 ~~16~~ 17 ~~18~~ 19 ~~20~~ 21 ~~22~~ 23 ~~24~~ 25 ....

We underline 3, then cross out every uncrossed third number after it, to produce

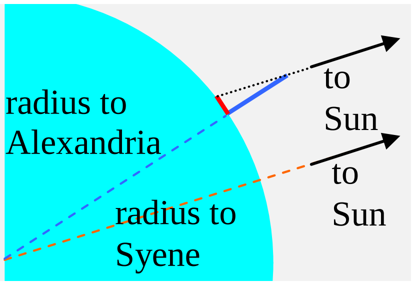
$\underline{2}$ $\underline{3}$ ~~4~~ 5 ~~6~~ 7 ~~8~~ ~~9~~ ~~10~~ 11 ~~12~~ 13 ~~14~~ ~~15~~ ~~16~~ 17 ~~18~~ 19 ~~20~~ ~~21~~ ~~22~~ 23 ~~24~~ 25 ....

Next is

$\underline{2}$ $\underline{3}$ ~~4~~ $\underline{5}$ ~~6~~ 7 ~~8~~ ~~9~~ ~~10~~ 11 ~~12~~ 13 ~~14~~ ~~15~~ ~~16~~ 17 ~~18~~ 19 ~~20~~ ~~21~~ ~~22~~ 23 ~~24~~ ~~25~~ ....

At each stage, the next uncrossed number is not divisible by any of its predecessors; therefore it is prime. We underline the number and cross out its multiples. Thus, the process underlines the primes and cancels all the composites, up until the end of the list. Such discoveries earned Eratosthenes, like Euclid, an invitation to Alexandria. There he became head of the Library and made a remarkable observation.

The Tropic of Cancer is the circle of Earth latitude along which the midday Sun is directly overhead on the longest northern day. In our era, the latitude is around 23.4°. (The circle of latitude does not go anywhere, but the Tropic moves, for a reason we will meet in the next section.) The Tropic now crosses Egypt at Lake Nasser, upriver from the dam at Aswan. That last was the site of the ancient city of Syene. Syene was widely known because on that longest day, the midday Sun shone right to the bottom of the deepest wells. In other words, Sun passed dead overhead. Eratosthenes observed that on the same day, the midday Sun at Alexandria cast shadows; it was not at the zenith. He correctly ascribed the difference to the curvature of Earth.



At right, we see the blue profile of Earth and two of its radii (dashed). The orange one reaches the surface at Syene. Its extension beyond the surface is the vertical there, and points directly at the overhead Sun. The blue one reaches Alexandria, where the Sun is in the same direction as at Syene, because *the Sun is infinitely distant*. That was a key assumption of Eratosthenes. It accords well with Sun's distance being so much greater than any Earth-bound distance. It does not harmonize with Sun's having positive size, but never mind. The extension of Alexandria's radius is a column (blue) standing vertically in the city. The column casts a shadow, shown red. From measuring the shadow and the column, Eratosthenes estimated the angle between the vertical and the Sun's direction at 1/50 of a circle (7.2° for us). That angle matches the central angle between the two radii (by alternate interior angles). Accordingly, the arc from Syene to Alexandria is 1/50 of Earth's circumference. Knowing that Syene was about 5000 stade (500 miles for us) from Alexandria—plus pretending that the distance goes due north, which is needed for the arc between them in our figure to have that length—he estimated the circumference of Earth to be 50 × 500 miles. That is a remarkable approximation of the 24,860 mi we accept as the "polar" circumference. (Earth is "oblate": bigger around the Equator.)

3. Hipparchus

Hipparchus was born around 190 BCE, near the deaths of Eratosthenes and Apollonius. He is in different contexts called "the father of astronomy" and "the father of trigonometry." He did remarkable analysis of the motions of Sun and Moon, refining and extending results of the Babylonians, including the estimate of the year.

Hipparchus introduced degree measure. The Babylonians had divided the chord of 60° into 60 equal parts. Using the chord is natural, given the ancient way of doing trigonometry; and making it 60 parts is natural, given their sexagesimal numeration. But notice that dividing the chord does not divide the angle into 60 equal parts. (See Exercise 3. The angle need not be central in a unit circle. In the exercise, we have it as one angle in a unit equilateral triangle with the opposite side marked off into 60 parts.) Hipparchus divided the *arc* of 60° into 60 equal parts. Or else he divided the entire circle into 360 equal parts. It is not known [to me, anyway] which of those choices motivated the definition of degrees. The 360 version corresponds roughly to the Sun's angular movement in a day. Hipparchus knew to within minutes that the year spans (365 + fraction) days. A non-integer is bad enough, but the closest integer is ugly too: 365 is divisible only by 5 and 73. On the other hand, 360 is divisible by 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 18 and their eleven partners (360/18 = 20, ...). For that reason, splitting the circle into 360 equal parts is a pragmatic choice.

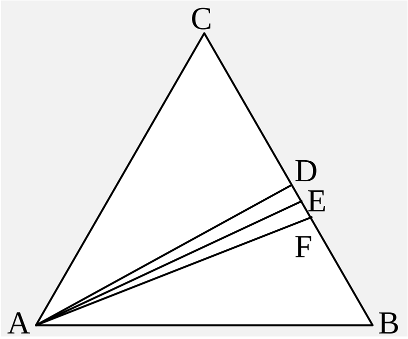
Among the angles Hipparchus measured were angles in the sky. The Babylonians had produced a coordinate system for star positions, measuring a sort of latitude north or south of the ecliptic and a kind of longitude eastward from the vernal equinox. Hipparchus performed new measurements. He found that the stars had moved toward the east. Remember that we said the stars do not move relative to one another (at least not enough for human eyes to detect, even over hundreds of years). What he found was that the entire dome of heaven had rotated eastward. He correctly ascribed the change to the *equinox's moving to the west*. His measurements allowed him to quantify the "precession of the equinoxes."

Earth’s rotation axis is not stable. Instead, it undergoes “precession,” the sort of wobble a top or toy gyroscope shows as it slows down. (Go to Wikipedia® for a great animation of precession.) The extension of the axis traces a circle in the sky over a period of around 26000 years. Since the axis is thus dancing, the Celestial Equator necessarily turns with it. Therefore the equinoxes (where the Equator meets the ecliptic) slide westward over the ecliptic (as Aristarchus suspected). In Babylonian times, the spring equinox was in the constellation Taurus. That put the start of spring and the long days in what we would call May. In our era, that equinox is two constellations to the west, in Pisces. Accordingly, our spring starts in March.

One consequence of the discovery was that two possible definitions of the year are inequivalent. We have described the year as the length of Sun’s trip around the starry dome. That is now called the “sidereal” year (year of the stars). We also described it as the length of the cycle of the seasons. That would be Sun’s trip from one spring equinox to the next. It is called the “tropical” year. The latter is shorter. In the time the Sun travels eastward from this year’s spring equinox toward next year’s, the equinox moves westward about 1/72 of a degree to meet it. (In the time of Hipparchus, his estimate was 1/100.) Therefore the tropical year is short of a sidereal year by about 1/72 of a day. (How long is that?)

Exercises III.C.3

1. a) Truncate the sieve of Eratosthenes to list the primes between 210 and 250. (Two hints: First, factor 210; second, if a number in that range is composite, then it has a prime factor less than $\sqrt{250}$.)
   b) Return to our illustration of the sieve. In the next stage, we underline 7 and cross out its multiples. What multiple of 7 is the first one that has not previously been crossed out?
2. (Boyer) Hipparchus knew from eclipse observations that as seen from the Moon, Earth has an angular size of about 2°, four times Luna’s size as seen from Earth. What Earth-Luna distance does that imply, assuming Eratosthenes was right (that Earth is a sphere of circumference 25,000 miles)?
3. In triangle ABC in the figure at right, each side has length 1. Put 59 equally spaced points along BC, to partition it into 60 equal parts. Let D be the 30th point, so that D is the midpoint of BC. Then let E be the 31st, F the 32nd.
   a) Show that tan(angle DAE) = $1/(30\sqrt{3})$.
   b) Show that tan(angle DAF) = 2 tan(angle DAE).
   c) Use trigonometry to prove that if $0 < \theta < 45°$, then
      $\tan 2\theta > 2 \tan \theta$.
   d) Show that angle DAE > angle EAF.
   e) (Calculus) Extend (c): Use the (extended) mean-value theorem to prove that if $k > 1$ and $0 < k\theta < \pi/2$, then $\tan k\theta > k \tan \theta$. (The tangent function grows faster than linearly. In simpler words: The tangent grows faster than the angle.)



4. Claudius Ptolemy

a) the solar system

Last of the great ancient astronomers was Claudius Ptolemy, born around 90 CE. He was from the family that ruled Greek Egypt; it is not surprising that he spent much of his life in and near Alexandria.

In one sense, he refined the work of Hipparchus. Claudius worked with degrees, and produced trigonometric tables for chords of degree-measured angles. It was Claudius who divided the degree into

Babylonian fractions. He divided one degree into 60 parts, each called *parte minuta prima*, first small part. The name **minute** has stuck, and indeed the mark used to denote minutes is called a "prime." (For example, we would write that Sun's angular size is about 32´.) He broke minutes into parts called *parte minuta secunda*, and you can see the origin of **seconds**.

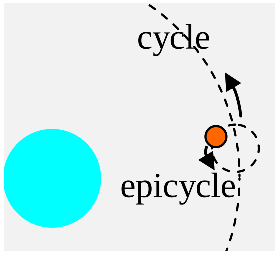
He applied his geometry to angle-measured organization of land and sky. On Earth's surface, he described a system of longitude and latitude for the Mediterranean world he knew. For the stars, he devised updated star charts. He organized the ecliptic into twelve zones spanning 30° each. (There you have 12 again.) To each of those, he assigned a constellation of the Zodiac. That assignment is strictly conventional: The actual star groups' expanses vary from the two Fishes, which together span 45° of the ecliptic (and 30° north to south), to the Crab, which spans barely 15°. The assignment required the creation (or at least completion) of Libra. Basically, Libra consists of two stars that had been the claws of the Scorpion.

[Much of today's *astrology* owes to Ptolemy's organizational scheme. That includes associating one-month intervals with the constellations the Sun would be traversing during those months in his era. Thus, the month starting with the spring equinox is matched with the Ram. Unfortunately, from Ptolemy's time to ours, precession moved the equinox west to the Fishes. As a result, the Sun now does not even *enter* the Ram until mid-April.

In case you're wondering, I'm a Taurus. The most important trait we Bulls share is our lack of superstition.]

Ptolemy's *magnum opus* is a book called *Almagest*. (Why would an ethnic Greek with a Roman name write a book with an obviously Arabic title?) In it, he proposed antiquity's most accurate model for the motions of the heavenly bodies. The motions needed explaining because they are not uniform. The Sun's motion is close to uniform. It moves along the ecliptic, toward the east, at the nearly constant rate of 360° per 365+ days. Not so the planets. Saturn—easiest to track because it is the slowest wanderer—spent the second half of 2011 moving east (the "normal" way) toward the line between the bright stars Arcturus and Spica. It crossed their line near end-year. In February 2012, it turned and headed *west* ("backward") toward the line, crossing it in May 2012. After June, it turned again eastward and crossed the line a third time. The "retrograde" motion, interrupting intervals of the prevailing eastward travel, bothered the ancients. They wanted to describe the wandering in terms of uniform motions.

Eudoxus had proposed that the wanderers are fixed to some transparent ("crystalline") spheres with distinct axes and (uniform) rotation rates. Apollonius dropped the spheres and described the *paths* of the planets with the "epicycle model." At right, we see the blue Earth at the center of a dashed circle. That circle is the orange planet's "cycle." The smaller dashed circle is the planet's "epicycle." The planet revolves counterclockwise at a fixed rate around the epicycle. At the same time, the center of the epicycle moves counterclockwise at constant rate around the cycle. If you tune the two sizes and rates the right way, then an observer on Earth sees the planet moving counterclockwise, except during the part of the epicycle closest to Earth. This description fit the ideal, of using uniform motions along a combination of circles. Unfortunately it was a poor fit to the data, the available observations of the planets. Others then added such modifications as having the center of the cycle itself orbiting a circle. Until Claudius, the most important ones came from Hipparchus and his measurements. However, it was Claudius Ptolemy's refinements that provided the most successful model, so much so that his projections of planetary motion were in use until telescopes were turned to the sky some 1400 years later.



[You must see the University of Nebraska's animation of the epicycle model. The animation incorporates Ptolemy's refinements, shows why the model explains the prevailing eastward travel interrupted by retrograde motion, and actually lets you "tune" the sizes of the circles and the rates of revolution.]

Exercises III.C.4a

1. a) Nowadays, the Tropic of Cancer is at north latitude 23.4° and Alexandria is (always was) at 31.2°. How many minutes of arc is that difference?
   b) By definition, a "nautical mile" of distance is what a minute of Earth latitude covers. One nautical mile equals 1.15 land ("statute") miles. What estimate does that imply for the circumference of Earth?

b) other science

(i) Claudius and the gap

It is remarkable how Greek astronomy exemplified the scientific method [which I long assumed to have begun with Galileo]. All these astronomers (you can add Archimedes) put together others' and their own observations, and especially their own *measurements*, to create geometric models of the heavens. The models allowed inferences, which in turn you could use improved measurements to refine.

Their studies stand out all the more against the gap of a thousand years after Claudius. During that interval, essentially no scientific inquiry and little mathematical study took place in the Roman and then Christian worlds. The absence is not a coincidence. Both Rome and the Church to which it gave birth were inimical to curiosity, the driver of mathematical and scientific thought. Rome was inclined to strictly practical knowledge, the Church emphasized the next life, and both highly valued deference to authority. Indeed, the centers of scientific inquiry moved about in response to social conditions. By the time of Alexander, it had already become hard for the science-minded to make a living in Athens. Moving the center of study to Alexandria was a natural. Over centuries, the center moved to Constantinople, then Baghdad, and only about 600 years ago to western Europe.

You should read Timothy Ferris's *Coming of Age in the Milky Way*. It covers all the astronomy above, and is especially valuable for depicting the Ptolemaic model as genuine science. Many of us are steeped in the idea that Claudius's solar system is just another unsophisticated attempt by ancient people to explain the phenomena of the world. [That idea was part of my education.] **Ferris** dispels that notion as coherently as the notion that people of Columbus's time thought that the world is flat.

(ii) Claudius and the mariner

There is a story [that I heard from Prof. Akin] connecting Christopher Columbus and Claudius.

Ptolemy's writings covered all of science, much the way Euclid's covered mathematics. Among his famous books was *Geography*. He saw Earth as smaller than it is. We referred to his system of latitude and longitude. Ptolemy estimated that Asia stretched to 180° in longitude, halfway around Earth. That happens to be true for northern Siberia, but not nearly for China and the Spice Islands (Indonesia). He also estimated Earth's circumference at 18,000 miles. The estimate used the distance from Rhodes to Alexandria, a baseline that is reasonable but hard to measure reliably; unlike the trip from Syene to Alexandria, it is over water.

That 18,000, though, was fine with Columbus. He was a master mariner. He would have known, for example, that if you sail *south* from Europe, below the Canary Islands, then the tropical winds and the corresponding currents will carry you west. Coming back, if you sail first north, then you get into winds and currents—what we now call the Gulf Stream—to speed you east. Columbus longed for the glory and wealth that would go to a man who sailed westward from Europe to the Indies. To mount such an expedition, he needed backing. To get it, he had to convince potential sponsors, as **Ferris** puts it, that the world is *small*. So, he lowballed even Ptolemy's low estimate, stretched Marco Polo's claims about how far Marco had gone overland to China, and claimed that the trip sailing west was under 4,000 miles. He proposed it to the Portuguese court. Lisbon, still possessed of its Moors and Jews, was the Alexandria of European science. The king consulted his geographers and astronomers. They told him Columbus was crazy: The Spice Islands were on the opposite side of a sphere 25,000 miles around; the trip would take three times as long as the food and water could hold out; to back this mad venture would be to toss men and money into the sea. The Portuguese passed, and Columbus went to the rubes in Castilla. There he found a teenaged queen, who persuaded her dim husband to finance the plan. The upshot, of course, was that the bumpkins got rich and the smart guys were late to begin exploration of the Americas.

The astronomers delivered one more opinion. Even with Columbus claiming to have reached the Indies, they insisted he had been maybe an eighth of the way around the globe, and proposed to give evidence. There was a total lunar eclipse predicted for Europe in the night of February 29, 1504. They predicted that it would start, in whatever land Columbus had actually reached, around sunset. Months before that day, Columbus had been forced to beach his storm- and worm-damaged ships on Jamaica. The aboriginal Arawaks had turned, with good reason, against Columbus's reprehensible crew, and the visitors faced starvation. Columbus decided to turn to heaven for help. He told the people he still called "indios" that his god was angry at them for their evil treatment of the whites and would destroy the world, beginning that night with the Moon. Sure enough, sundown revealed a Moon with part missing. More and more darkened out with the passing minutes. The Arawaks wailed and prayed, to no avail. Finally, they asked Columbus (as his son wrote) to intercede with his god in their behalf. Columbus promised to try, retired to his quarters, and got the Moon back. The natives resumed providing the supplies that allowed Columbus and his men to survive until the Spaniards sent rescue vessels.

5. The Calendar

Our account has now passed irreversibly into the Common Era. It is therefore worthwhile to discuss how the numbering system for our years came to be. It is also fitting that we cover Rome, since so much of our culture is based on hers, including the years' numbering. Separately, it is apt that we limit the coverage to an afterthought like this. In the development of science and mathematics, Rome was at best useless, at worst destructive. That seems impossible, given the Romans' skill and achievements in architecture, civil engineering, communications, even warfare. Still, it seems they had no use for intellectual pursuits that did not immediately produce buildings, roads, food and water, or dominions.

a) the months

Tradition has it that Rome was founded in 753 BCE. It also says that one of the founders was suckled by a wolf; suspend disbelief. Fairly early, Romans accepted months named Martius, after the god of war; Aprilis, perhaps referring to the opening of the flowers; Maius, after either Mercury's mother or the goddess of majesty; Juno, after Mrs. Jupiter; and Quintilius, Sextilius, September through December, equivalent to Month5 through Month10. At some point, they added Januarius, named for the god of beginnings, and Februarius (for a festival?).

By 500, they had dropped their monarchy and replaced it with a republic. The republic's top officials were magistrates called "consuls." They were elected in December and began their terms at its end. The consul system had such popular support that people began to think of the year as concurrent with the consuls' terms. They started thinking of Januarius as the first month. Notice the strange effect: The months *named* 5-10 became numbers 7-12, so that **Dec**ember is still 12.

Lunar tradition still applied; the twelve months spanned just 354 days. (What does 354 have to do with the Moon?) Consequently it was necessary to “intercalate,” to stick in an extra month a little more often than once every three years. (A 354-day calendar leaves out 11¼ × 3 extra days every three cycles of the seasons. The Muslim calendar is that long and strictly lunar. It intercalates days—something like 8 every 33 years—to synchronize with the *Moon*, not months to sync with the Sun. Accordingly, its holy days retreat 11 or so days per year: Thus, Ramadan started August 11 2010, August 1 2011, July 20 2012.) The intercalation was left to the *decemviri*, a council of ten men chosen no doubt for loyalty, as opposed to intelligence or honesty. Their intercalations were so variable that by the middle of the first century BCE, the beginning of the year (officially Martius) was almost a complete season off. It was then that the Senate decided to hand the fate of Rome to a member of the Julii.

b) the leaps

Gaius Julius Caesar was already an important man in 49 BCE. He was a senator and former consul, he was rich [But I repeat myself.], and he was *pontifex maximus*, chief of the bridge makers. The *pontifices* made and cared for the bridges over the Tiber; Julius was in effect Defender of Rome. In 49, the Senate voted to make him dictator for ten years. Notice, his word would be law, but for a fixed term, without inheritance.

Julius was a pragmatic man. (What did he do for a living?) He wanted Rome to have a fixed calendar, one by which a grocer in Rome, a farmer in Tuscany, and an army captain in England would all know today’s date in the capital. Calendars being products of astronomy, he summoned the leading astronomer of Alexandria, Sosigenes the Greek. (How did Julius know about Alexandria?) Doubtless Sosigenes observed that Rome’s empire, importing food and exporting soldiers from a port that froze in winter—Rome is at the latitude of New York—needed to track the seasons. The cycle of the seasons covers 365¼ days, just about. A calendar of 365 days, with an intercalated day every four years, would track the seasons to within a day in about a century (Exercise 1). Presumably as a result, Caesar said, listen up Rome, we are going to do as follows. First, we are going to delay 45 BCE for eighty days, to put the beginning of spring back at the start of the calendar. Second, that year and every fourth year will have 366 days, consisting of alternating months of 31 and 30 days. Finally, the intervening years will have 365 days, with the adjustments made where we always make them, in February. [I did lie about “listen up” and “45 BCE.” You know there are untruths here, as in the previous clause.] That ruling established this sequence of months:

Martius	31 days
Aprilis	30
Maius	31
Juno	30
Quintilius	31
Sextilius	30
September	31
October	30
November	31
December	30
Januarius	31
Februarius	30 or 29

A year later, a group of senators voted Julius out. The senators that did not participate in the assassination renamed Quintilius “Julius.” Two years after that, the *decemviri* interpreted “every fourth year” in a Roman way: leap, two, three, leap, two, three, leap .... [I once heard Roman counting described this way: Imagine a driving course along which there is a marker at the start, another at the one-mile mark, another at two miles, another at three miles; the Romans would count, not three miles, but four markers.] They put a leap into what we would call every third year.

It took 13 years to settle the leadership war between Caesar's closest lieutenant, Marcus Antonius, and his nephew and adopted son, Octavian. Finally in 31, the forces of Octavian and Agrippa defeated those of Marcus and the last of the Greek rulers of Egypt. (What was the name of that last ruling Ptolemy?). Octavian became undisputed leader of Rome. Four years later, the Senate proclaimed him "Augustus," first *imperator* of Rome. He, who was a decent ruler for more than 40 years, finally settled the leaps. In 9 BCE, after 12 leaps in 36 years—three too many—he ruled that Rome would skip the next three leaps. In 8 CE, he resumed the leaps. Augustus had set the Sun back into harmony with Rome. The Senate voted to rename Sextilius "Augustus." You can see the problem with that: Julius was longer. So 31 days were assigned to Augustus. An adjustment was needed somewhere; Februarius was elected, changing to 29 or 28 days. Finally, the quarter Julius-September was excessively long, at 93 days. The days were redistributed to 30 in September, 31 in October, 30 in November, 31 in December. Thus, barely over 2000 years ago, the calendar familiar to us was in place in Rome.

c) the eras

Sixteen centuries after Augustus, Rome still had an empire. Its leader was still called *pontifex maximus*, "Supreme Pontiff." Those men were sentimental about the spring equinox, because they believed it pointed to the resurrection of the ruler of the universe. We therefore turn our gaze to a Jewish boy born to the name Joshua ben Joseph.

There are no records from the lifetime of the man the Romans called Jesus the Nazarene. The biographies of Jesus, the Gospels, were written scores of years after his death. They do agree about the circumstances of his death. They say that he was arrested after a *seder*, the *Pesach* (Passover) dinner; that he was tried before Pontius Pilate and put to the cross the next day; and that the day was day before *shabbos* (Sabbath). For Pontius, there are plenty of records. He governed Judea from 26 (12 years past the death of Augustus) to 36 CE.

That Gospel description carries astronomical information. It says that Jesus was taken on Passover, a Jewish holy night marked by the first full Moon following the spring equinox, and that the next day was Friday. Evidently, that full Moon does not happen on Thursday every year. During Pontius's tenure, it happened in two years, 30 and 33. It is safe to conclude that Jesus died one of those two years.

Within 100 years, Christianity became an important religion in Rome. Within 300, it became dominant, so much so that the emperor Constantine converted. He moved the capital to Byzantium, for which he found the convenient new name "Constantinople." In 325, he convened the Council of Nicæa (birthplace of Hipparchus), which was a kind of constitutional convention for the Christian Church. The Council set out many of the Church's tenets. It also began a search for information about Jesus. The search had an important result two centuries later, when the scholar Dionysius Exiguus ("Little Dennis") concluded that Jesus was born in the 28th year of Augustus. That year was 525 years before (Dennis's work). He named it "1 *Anno Domini*," first year (in the era) "of our Lord." He called the year before that "first year before our Lord." That was the origin of the designation of the eras as AD and BC.

> Dennis's numbering amounted to beginning a new era at year +1 and calling the preceding year -1. Why was there no year number 0?
>
> There is wide disagreement as to which year Jesus was actually born, but 4 BCE is a good candidate for several reasons. Although 1 CE was the 28th year of "Augustus," that man was governing Rome for four years before gaining the name. Putting the birth of Jesus in 4 BCE and his death in 30 CE agrees with the accounts that say he was in his 34th year at his crucifixion.

The numbering system did not come into wide use, even in the Church, for hundreds of years. The convention to use “Common Era” and “Before Common Era,” removing the religious content in AD and BC, is only about 40 years old. [It is a mystery to me why English used the Latin “Anno Domini” into the 20th century, but not the Latin equivalent for “Before Christ.”]

d) the skips

Dennis made one other inference. He estimated that when the Council met in 325, the spring equinox happened on March 21. The Church accordingly declared that Easter would be celebrated on the Sunday following the (“Paschal,” maybe rooted in *Pesach*) first full Moon after March 21.

Recall that the Julian calendar is too long, causing the seasons to retreat through the calendar at the rate of roughly one day every 130 years (Exercise 2a). In the 1250 years after 325, the spring equinox fell back to around March 11. Anchoring Easter, a holiday the New Testament tied closely to the time of flowering, to March 21 threatened to push it into the summer.

> Already around 700, the Venerable Bede—an early advocate for Dennis’ AD/BC scheme—was warning about this date drift and calling for adjustment to the calendar.
>
> By 1513, Juan Ponce de León had worn out his welcome as governor of the island he had named “Puerto Rico.” (Columbus had called it “San Juan,” still the name of its capital.) Ponce asked the king of Spain for permission to go colonize Bimini, in the Bahamas. Permission given, he set off northwest. He was as good at navigation as at politics, so he missed. He ended up on North America, arriving on Easter. Since Spanish calls the Christmas and Easter holidays *pascuas natales* and *pascuas floridas* (the birth celebration and the flowering one), he named the place where he landed “la Florida” (flo-REE-dah), claimed it for Spain, and invented the Early-Bird Special.

In 1576, Pope Gregory XIII [unsuperstitious, like us Taureans] decided to make the adjustment. (Why then? How had the Church changed since, say, 1350, when the calendar drift was almost equally obvious?) He constituted a committee, which decided to follow recommendations made by the late Luigi Lilio (sometimes written “Giglio”). Lilio had suggested bringing the equinox back to March 21 by *jumping* ten dates, and keeping it there by removing three leaps every 400 years. He even suggested which three to skip. He said skip, for convenience, the century years that are not divisible by 400. Gregory ruled accordingly, decreeing that the day following October 4 1582 would have the date October 15 1582; and that the years 1600, 2000, 2400, ..., would be leap years, but not 1700, 1800, 1900, nor 2100, 2200, ... That is how our yearly calendar is now set up.

> The “Gregorian Calendar” is now nearly universal, but like Dennis’s system, it was not accepted immediately. England was ruled in 1582 by a daughter of Henry VIII. Henry had fallen out with the Church, because it wanted him to buy his wives, and Henry preferred to lease. Therefore Gregory’s word was not law in Protestant England. She did not adopt the calendar until 1750. (Similar question: Why then? How was England different in 1750 from 1582?) Likewise, Orthodox Greece and Russia resisted into the twentieth century, respectively 1922 and 1918. (Why was Russia different in 1918 from, say, 1910?)
>
> [Remember the Y2K bug? My computer will not let me set the date to year 2100. I don’t know for sure, but maybe its programming is tuned to our age, meaning 1901-2099. During that span, *every* fourth year is leap. Do you suppose there will be a Y2.1K bug?]

Exercises III.C.5

1. Sosigenes would have known Hipparchus's estimate of the tropical year,
   365 + 14/60 + 21/3600 days.
   (Why would he have known it?) What is the reciprocal of the difference between that number and 365¼? (Sosigenes must have told Julius that a calendar averaging 365¼ days would need adjustment by one day in that many years.)
2. The current estimate of the tropical year is 365.242374 days. (Precession is periodic, and therefore so is this estimate.)
   a) What is the reciprocal of the difference between that number and 365¼? That reciprocal is the number of years in which the Julian calendar advances one day relative to the cycle of the seasons.
   b) The Gregorian calendar has 365 days plus 97 intercalated days every 400 years. Using the given length for the tropical year, in what year will it have advanced one day?
   c) According to (b), the Gregorian year is 365 + 97/400 days. Where did we encounter that number before?
3. [On February 29 1996, I received a paycheck. I did not remember ever before getting paid on "leap day," so I wondered when it would happen next. City College has its regular paycheck issuance every other Thursday. When will it next have a payday on February 29?]
4. a) Check that the Gregorian rule provides the following: Every 4th year is leap, except that you skip every 25th leap, but then you re-leap every 4th skip.
   b) Suppose you extend that provision *ad infinitum*: Leap every 4th year, skip every 25th leap, re-leap every 4th skip; re-skip every 25th re-leap, ... How long will the resulting year be?

Chapter IV. Middle Peoples

India and China have long mathematical traditions. Indeed, they had robust and highly cultured civilizations when Europeans were barely out of the hunter-gatherer stage. In this chapter, we will see how Chinese and Indian discoveries were well ahead of the European, and take a look at the Americas.

Section IV.A. India

Alexander conquered western India, all the way through the valley of the Indus. (Remember that modern-day Pakistan came into existence when Gandhi's dream fell apart in 1947.) The campaign brought influence from the west to Indian mathematics. Oddly, it was not Greek influence. What the Macedonian brought was from Mesopotamia, **Struik**'s "Oriental tradition." Consequently, geometry was not a great interest, and none of Indian mathematics reflected a deductive approach. Instead, the focus was numerical. The development was intuitive and the works were prescriptive, exhibiting instructions, methods, and the like.

1. Geometry

The earliest Indian books were the *Sulvasutras*, literally "books of rules about cords." The name suggests association with surveying, with good reason. The books were compendia of geometric information, on such topics as Pythagorean triples and measure formulas.

Some of their mensuration was mistaken. The errors were not simply approximations, as in the Egyptian rule that a circle has the area of a square 8/9 as wide (section III.A.1). They were actually erroneous statements, as in claiming that the area of a quadrilateral is the product of the averages of opposite sides.

A better example came from the extensive work of Brahmagupta. The Indians knew the equivalent of **Heron's formula**: A triangle of sides a, b, c, with semiperimeter $s = (a + b + c)/2$, has area

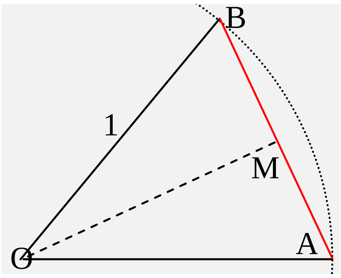
$$A = \sqrt{(s[s-a][s-b][s-c])}.$$

(An elementary, but necessarily complicated, geometric proof is at the University of Georgia.) Brahmagupta (6th century CE) gave a generalization for a quadrilateral with sides a, b, c, d:

$$A = \sqrt{(s[s-a][s-b][s-c][s-d])},$$

with s the new semiperimeter. Bhaskara (12th century) in the *Lilavati* observed that Brahmagupta's formula could not be right, because a quadrilateral is not determined by its four sides. That is, two triangles of matching sides must be of equal size, by SSS; but a square and non-rectangular rhombus of equal sides do not have the same area. The observation was perceptive, but Bhaskara seems to have been unaware that Brahmagupta's formula works for a quadrilateral inscribable in a circle. (Compare Exercise III.A.3:6. See also **Boyer** page 242. At page 233, Prof. Boyer recalls how the Arabic philosopher Muhammad al-Biruni (973-1048), commenting on the odd combination of truth and error in Indian mathematics, described it as a mixture of "common pebbles and costly crystals.")

For us, it is Indian trigonometry that is important. It was the first to look like ours. The *Surya Siddhanta* ("Sun system," with obvious astronomical connection and going back possibly to BCE) introduced the half-chord. Recall that Greek trigonometry worked with the chord AB (red at right) of central angle AOB. The Indians drew the angle bisector OM (dashed), which necessarily bisects the chord, and called AM the **half-chord** of angle AOM. (Compare section III.A.8a plus Exercise 1 there.) That length is precisely our definition of the sine of the angle. Our cosine, tangent, and the other functions are in that *Siddhanta* as well.



2. Numeration and Arithmetic

The other way in which Indian mathematics resembled ours—meaning, of course, in which ours follows the Indian—is “ciphered positional decimal” numeration. *Decimal* numeration is ancient. The Egyptians used decimal aggregates (symbols for 10 and its powers). So did the Romans, with the added convenience of symbols for the 5-multiples 5, 50, and 500. *Positional*, or *place-value*, numeration was the Babylonian way. In geometry and astronomy, their sexagesimal numeration persisted until just centuries ago. *Ciphered* numeration, using symbols instead of marks or aggregates for the digits, also predates the Indians. (Here, “cipher” has nothing to do with codes. The Romance languages, and some others, use words related to it for “digit.”) The Egyptians had some cipherization as far back as 2000 BCE, and the Greeks used letters of the alphabet to represent some numbers. All those elements had coalesced in India by 600 CE.

The Indians came to represent any natural number by a string of symbols chosen from nine that have evolved into our

1 2 3 4 5 6 7 8 9;

see **Boyer**. What we call “Arabic numerals” are Indian numerals.

The lack of a zero symbol made for the usual difficulty. When a power of 10 was missing, a space or other indication was needed. Recall that the Babylonians faced the same problem, and did not create any zero symbol until very late. The culmination of Indian numeration was the adoption, before 876, of the round symbol we now use for zero.

We have remarked that in aggregate-based systems, addition is easy and multiplication hard. It is worthwhile to look at base-based algorithms for arithmetic.

> Consider the multiplication process in the array at right. It uses the distributive law,
>
> $$\begin{aligned} 567 \times 89 &= (7 + 60 + 500)(9 + 80) \\ &= 7 \times 9 + 60 \times 9 + 500 \times 9 + 7 \times 80 + 60 \times 80 + 500 \times 80. \end{aligned}$$
>
> However, it multiplies only digits. The 7 × 9 = 63 is first. For 60 × 9, we need just 6 × 9 = 54, offset so that its value is 54 tens. The offset is why in each pink square, we may write the zero, or instead leave the square empty. The array then displays the remaining digit multiplications, all blue.
>
> Does the array look unfamiliar? We normally “carry” from place to place. Here it is clear that carrying simply saves space. It holds the multiplication by each digit in 89 to a single line. On the other hand, operating right-to-left agrees with our usual way. We see the advantage of doing so: It facilitates placing the digit-products in the right columns.
>
> The illustration continues with the sums (red) of the digit-products. If we insist on refusing to “carry,” we may again write multi-digit sums on separate lines, in the appropriate columns. We then keep adding until all the sums are less than 10. Computers do something like that, but in base 2.

		5	6	7
		x	8	9
			6	3
		5	4	▨
	4	5	▨	▨
		5	6	▨
	4	8	▨	▨
4	0	▨	▨	▨
				3
		1	6	▨
	2	3	▨	▨
4	8	▨	▨	▨
		4	6	3
1	0	▨	▨	▨
4	▨	▨	▨	▨
5	0	4	6	3

There is another algorithm, although in use it has become extinct. Look at the following approximation of $\sqrt{567.8}$. (We allow ourselves decimal fractions; the Indians never wrote them.) To begin, it *pairs* the digits leftward and rightward from the decimal point (left-hand panel in the figure on the next page). It supplies the 0 if needed, plus as many pairs 00 as the answer’s desired number of decimal places calls for. It then iterates the following set of instructions:

1. Create a number on the red line by

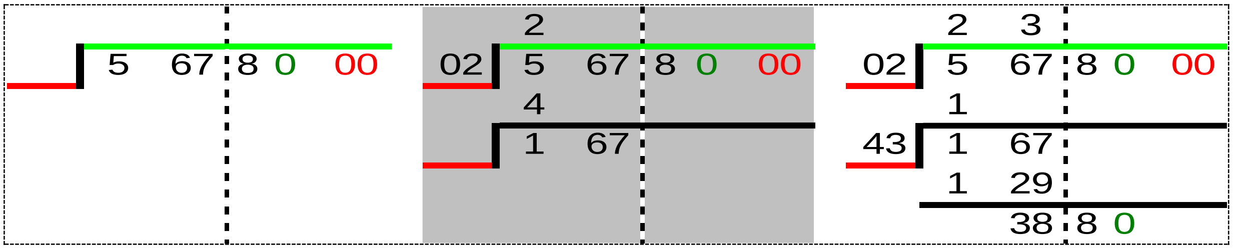
a) doubling what is currently on the green line and

b) placing a digit at the end (units place) of the red line and above the current pair. The same digit goes at both places. Make the digit the largest such that *it* times the created number does not exceed the current “dividend.”

2. Multiply digit by created number, then subtract from dividend.

3. Bring down the next pair to create the next dividend.

4. Repeat as necessary, for the number of decimal places desired in the approximation.



We want to approximate $\sqrt{567.8}$ to two places. In the left-hand panel of the box above, we supply “0 00” after the decimal point. On the red line, we (a) start with 0, from doubling the empty green, (b) then adjoin 2 (middle panel). That makes 02 the created number and 2 the digit. *We could fit 1 on the green line, because* $1 \times 01 \le 5$*; and 2, because* $2 \times 02 \le 5$*; but not 3, because* 3×03 *exceeds 5.*

So far, it appears that we have $u = 2$ as the largest digit whose square is 5 or less. In fact, what we found is the greatest 10-multiple $10u = 20$ such that $(10u)^2 \le 567.8$. This square falls short by 167.8. When we bring down the 67 pair, it is clear that we next work with just the 167 part of the shortfall.

To restart, (a) we double 2 to put 4 on the next red line (right-hand panel). Then (b) the digit is 3 and the created number 43. The reason is that $3 \times 43 \le 167$, but $4 \times 44 > 167$. Multiply, subtract, bring down, and the next dividend is 3880. During this stage, we found $v = 3$ to be the largest digit with

$$v(2u \times 10 + v) \quad \le \quad 167.$$

Do you see why that is related to squares? From

$$v(2u \times 10 + v) \quad \le \quad 167 \quad = \quad 567 - (10u)^2,$$

we have

$$(10u)^2 + 2u \times 10 \times v + v^2 \quad \le \quad 567.$$

We recognize $(10u + v)^2$ on the left. Our process so far has identified the biggest two-digit number $10u + v$ whose square is 567 or less.

That gives you an idea why the method works. Complete it in Exercise 2.

A few things are reasonably evident. First, if the square root is a terminating decimal, like $\sqrt{552.25} = 23.5$, then the algorithm will arrive at a zero dividend and terminate at the exact root. Second, it does not matter how many digits there are either side of the decimal point. If the original number had been 56,780, then the algorithm would still have produced 2. The 2 would have been one column further left, representing 200, the biggest multiple of 100 whose square is 50,000 or less. Then it would produce 3, because 230 is the biggest multiple of 10 squaring to less than 56,700. (Check!) Third, the multiplication algorithm and the square-root algorithm both take advantage of place-value numeration, not specifically *decimal* numeration. Try Exercises 1 and 3.

[Richard Anderson was a contributor to the mathematical community as well as to mathematics. Around 1985, I heard him address the American Association for the Advancement of Science on the subject of technology’s effect on the math curriculum. His teachers had learned the formula for solving cubic equations (which we will meet later), but did not teach it to him. They did teach him our square-root algorithm, but he had found it unnecessary to teach it to his students. He taught his students linear interpolation in the use of log and trig tables, but they did not need to teach it to their students.

Back then, the first graphing calculators were coming out. It fell to Anderson's grandstudents to consider the possibility that teaching approximation and equation-solving methods might be rendered obsolete. Now, hand-held devices can carry out symbolic algebra and calculus. If you are going to teach mathematics, expect parts of your learning to turn into quaint relics, and prepare to adapt.]

Exercises IV.A.2

1. Use a multiplication algorithm—either as above or in the familiar way—to multiply
   123(base 5) × 432(base 5).
   Check by turning the numbers into base 10.
2. Continue the square-root algorithm in this section to approximate √567.8 to two decimal places (truncated in the second place).
3. Use the square-root algorithm in base 5 to find the integer part of √120241(base 5). (The radicand is the answer in Exercise 1, and the square root is decimal 66+.)

3. Extension of Arithmetic

Brahmagupta's geometry was in the service of his astronomy, but he also contributed where algebra meets number theory. He defined the arithmetic of negative numbers six centuries before they were known in Europe, as well as the arithmetic of zero long before "0" entered the numeration.

The "natural numbers" really are natural, having quantitative significance that must have been in the minds of the earliest humans. We, dealing with (among other things) commerce and temperature, have no trouble using numbers below 1. Even early peoples, however, must have dealt with bodies of water whose levels rise and fall with spring melt, droughts, and the tides. In any of those situations, you can make one mark where the lowest water was, expecting to see the water rise some marks above the low point. If later the extreme low becomes lower, then you can talk about the level reaching so many marks *below* the original mark. That gives some *quantitative* meaning to "numbers" less than 1.

This *levels* interpretation, though, is not enough to qualify them as *numbers*. Let us agree that for things to be called that, there must be a way to operate on them, to do arithmetic. We will decide how to designate them and how to add, subtract, multiply, and divide.

> One choice is to call the number right before 1 "before1." Then the next number down would be "beforebefore1." Too clumsy—make that one 2before1, then continue 3before1, ... In that case, make the first one 1before1. Notice that this is exactly what Little Dennis did in numbering the years
>
> ... 2 BC, 1 BC, 1 AD, 2 AD, ...
>
> We can then easily define additions with these new things by continuing the pattern
>
> 1 + 3 = 4,
> 1 + 2 = 3,
> 1 + 1 = 2.
>
> The next left side is 1 + 1before1, and the next right side is 1. Therefore we continue
>
> 1 + 1before1 = 1,
> 1 + 2before1 = 1before1,
> 1 + 3before1 = 2before1, ...
>
> We can deduce this addition rule: If *m* and *n* are natural numbers, then

> $m + n$before1 = $m - n + 1$ if $m \geq n$,
> $m + n$before1 = $(n - m)$before1 if $m < n$.
>
> That need for a two-case rule is an annoyance. Dennis's system has the same irksome property. Augustus ruled Rome from 31 BCE until he died in 14 CE. He ruled for
>
> 14 CE – 31 BCE
>
> years, a difference that looks like 45 years. Subtraction is defined in terms of addition: By definition, that difference is what you would add to 31 BCE to make 14 CE. Since
>
> 44 + 31before1 = 44 – 31 + 1 = 14,
>
> we conclude that he ruled 44 years. We need to change the names of the numbers.

Call the number before one "zero." We are not saying, invent a symbol other than space to indicate a missing power of 10. We are saying, give a name to the *number* that signifies how many oranges are left if you start with one and lose it. Then invent *even lower* numbers: 1below, 2below, ..., the same names we attach to temperatures. Our pattern then reads

1 + 1 = 2,
1 + zero = 1,
1 + 1below = zero,
1 + 2below = 1below.

The addition rule becomes *one*, without separate cases: To add numbers on opposite sides of zero, subtract the absolute values and attach the "sign" that went with the higher value.

Let us adopt the familiar designations 0, -1, -2, ... By decreasing the first terms, we establish

2 + -1 = 1,
1 + -1 = 0,
0 + -1 = -1,
-1 + -1 = -2.

We see, for example, that to add numbers of like signs, we add values and replicate sign. Furthermore, adding 0 leaves the other summand unchanged.

Once we define addition, it defines subtraction. Thus, 3 – -5 is that number whose sum with -5 is 3. Since

8 + -5 = 3,

we see that subtracting gives the same result as sign-changing the subtrahend [or subtractee or whatever you call that second term] and adding. Observe then that, unlike with just the natural numbers, *all* integer subtractions are defined.

For multiplication, one combination is natural. If we think of multiplication as repeated addition, then we immediately have

-2 × 3 = -2 + -2 + -2,

and we have already agreed that the latter result is -6. Separately, we see that in

2 × 3 = 6,
1 × 3 = 3,

the products are decreasing by 3. Accordingly, the next three results are

0 × 3 = 0,
-1 × 3 = -3,
-2 × 3 = -6.

The patterns: Multiplication by 0 always gives 0, and the product of unlike signs is always negative.

The other combination, negative times negative, is the one that always makes trouble. We can follow two interpretations to resolve it. First, we thought of -2 × 3 as repeated addition. How about if we say, to multiply by -2, first change the other factor's sign, then multiply by 2? Then

$$-2 \times (-3) = 2 \times (3).$$

Alternatively, we can return to our patterns. In

$$-2 \times 2 = -4,$$
$$-2 \times 1 = -2,$$
$$-2 \times 0 = 0,$$

we see that the products are *increasing* by 2. We then have to continue with

$$-2 \times -1 = 2,$$
$$-2 \times -2 = 4.$$

We must accept that the product of like signs is positive.

Finally, division answers a multiplication question, and is therefore not always defined within the integers. We have $-12 \div 6 = -2$, *because we have agreed that* $6 \times -2 = -12$. However, there is no integer whose product with 6 is -13; $-13 \div 6$ is undefined. More important, division by 0 is undefined. To write $5 \div 0 = x$, you would need $5 = 0 \times x$; the latter is impossible. To write $0 \div 0 = y$, you would need $0 = 0 \times y$; not a problem, except that it does not *uniquely* specify y.

Exercises IV.A.3

1. a) Give a definition that "invents" the "rational number" -13/6.
   b) Use your definition to *prove* that -13/6 is negative.
   c) Use your definition to prove that $-3 < -13/6 < -2$.
2. Evaluate the sum
   $$1 - 2 + 3 - 4 + \ldots + 99 - 100.$$
   Mention the arithmetic or algebraic principles that allow you to proceed.

4. Algebra and Number Theory

Brahmagupta's algebra was remarkably advanced. It included extensive treatment of indeterminate equations, including certain kinds of quadratics. It displayed, for the first time, a willingness to allow negative roots. Here we highlight his complete solution of the linear Diophantine equation.

In section III.B.5, we looked at the general form

$$ax + by = c.$$

Here a, b, and c are fixed integers, and the hunt is for integer solutions x and y. We noted that we can limit attention to the case in which a and b are relatively prime. For the example

$$6x + -25y = 73,$$

we found one solution, $x = 1533$, $y = 365$.

> Now assume $x = s$, $y = t$ is another solution. We subtract the two equations
>
> $$6s + -25t = 73,$$
> $$6(1533) + -25(365) = 73,$$
>
> and transpose to write
>
> $$6(s - 1533) = 25(t - 365) = -25(365 - t).$$
>
> The left side is a multiple of 6, so 6 divides $-25(365 - t)$. Because 6 is relatively prime to -25, we conclude that 6 divides $(365 - t)$. (The theorem that if a number divides a product and is prime to one factor, then it must divide the other factor, is Exercise III.B.4a:5.) For the same reason, -25 divides $(s - 1533)$. In fact, rewriting
>
> $$(s - 1533)/-25 = (365 - t)/6,$$
>
> we see that $(s - 1533)$ is the same multiple of -25 that $(365 - t)$ is of 6. Call the multiplier m. We can

> then characterize the structure of every solution $x = s$, $y = t$:
> $s = 1533 + \text{-}25m$, $t = 365 - 6m$.
>
> In words, given the one solution $x = 1533$, $y = 365$, you get every other solution by adding to 1533 a multiple of -25 (in the general form, a multiple of b) and subtracting from 365 the like multiple of 6 (respectively, of a).

You can see that this argument is more number-theoretic than algebraic. The influences of Brahmagupta and others made number theory the focus of Indian mathematics through and far beyond the Middle Ages. The achievements in this area are noteworthy, but none more so than those just a century ago of Srinivasa Ramanujan (Rah-MAH-nu-jam). In a number of ways, his work was the epitome of Indian mathematics: given to numerical relations; intuitive rather than deductive, with results seeming to spring full-blown from his mind; occasionally false; brilliant.

Ramanujan was born in 1887. As boy and young man, he showed flashes of mathematical brilliance, but was not good enough overall to get a university degree. Around 1912, he sent stacks of results to some English mathematicians. (Many of the results involved infinite series. **Merzbach**, p. 202, indicates that series were known in India by the end of the 1300's.) They were roundly ignored, except by the renowned Godfrey Hardy. He alone recognized that they were not just largely correct, but actually works of genius. He got Ramanujan to Cambridge, where the son of tropical India continued a prodigious output until the English weather killed him in 1920. Read his story in this book: Robert Kanigel, *The Man Who Knew Infinity*.

[It is common for math departments to receive manuscripts from non-academics evincing earth-shaking discoveries. (They are never routine discoveries.) These include astonishing formulas, amazingly short proofs of either deep or unestablished results, or arguments showing that some famous theorem is false. To be fair to Hardy's colleagues, we have to admit that deep math from a clerk in the Indian port of Madras would have been as unlikely as important physics from a clerk in the Swiss patent office at Bern.]

Exercises IV.A.4

1. Describe all the integer solutions of the equation
   $56x + 30y = 16.$
2. In what way did the influence of Greek mathematics upon ours differ from Indian influence?

Section IV.B. The Middle Kingdom

Any coverage of China has to face ancient China's relative isolation. Trade and interaction certainly existed, especially with India, but Chinese discoveries only slowly drifted westward. China developed gunpowder and printing around the tenth century, paper and the magnetic compass by the eleventh, all of them novelties in Europe when the age of exploration began at the end of the fifteenth. Another factor is the orientation of this book. It intends to culminate with mathematical developments centered on nineteenth century Europe. For those, there was little Chinese influence. Further, the record of Chinese science has gaps. Their media were not as hardy as the Babylonians' baked clay tablets and were not favored by Egypt's dryness for preservation. Therefore even when we know Chinese discoveries, it is difficult to date them.

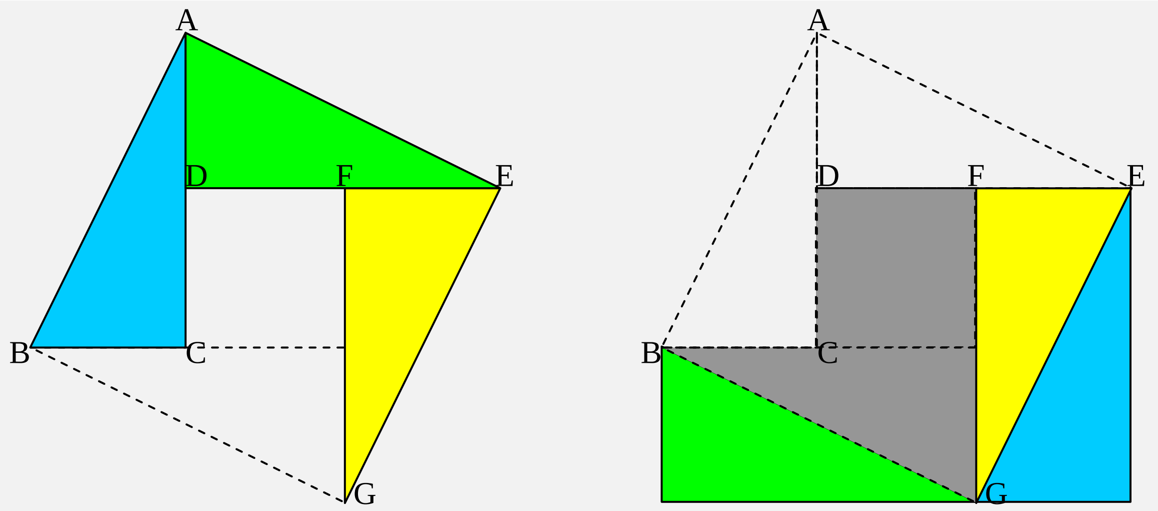
1. Geometry

Evidently, a people with advanced architecture, civil engineering, and cartography must have known sophisticated geometry. We do not have evidence of deductive development like that of the Greeks.

Indeed, there was no Greek influence in Chinese mathematics, since Alexander's conquests did not reach far enough. However, we do have evidence of parts of their knowledge.

Some material on the properties of right triangles is in the earliest known mathematical text, the *Arithmetical Classic of the Gnomon and the Circular Paths of Heaven*. It contains a picture that clearly serves as *proof* of the Pythagorean theorem. We cannot know whether it preceded the Greeks' proof, because the best we know of its date is that it traces to the Zhou Dynasty, 1046-256 BCE.

In the left panel of the figure below, we have right triangle ABC, shaded blue. We adjoin a copy (green) of it by going down from A to D with AD = BC, then right to E with DE = AC. Repeat (yellow) with EF = BC, FG = AC. At that stage, BC and its extension rightward and BG bound an uncolored right triangle whose legs must match AC and BC, making it also congruent to ABC.



We now conclude that ABGE is the square on the hypotenuse. Its four sides are congruent, same length as AB, and each of its four angles is the sum of the complementary acute angles of triangle ABC. (In the Chinese picture, this tilted square was inscribed within an upright square subdivided into seven rows and seven columns, clearly indicating that ABC was a 3-4-5 triangle. However, our construction is independent of the dimensions of triangle ABC.)

In the right-hand panel, we move the blue triangle to the bottom right-hand corner and the green triangle to the bottom left. We fill with gray the previously uncolored square and triangle. The stair-step region now having all the colors comprises the square on BC on the left, the square on DE = AC on the right. We have disassembled the square on AB, then reassembled the pieces into the squares on AC and BC. We have shown that the square on the hypotenuse is the sum of the other two.

That the picture proves the theorem is undeniable, but it yields a puzzle. The argument clearly depends on the statement that the right triangle's acute angles sum to a right angle. That statement is equivalent to the angle-sum theorem. Also, the original square's coming together and "moving" the triangles involve considerations of congruent triangles. But the geometry in the *Classic* does not include those properties of triangles; it does not display the development in Euclid's *Elements*.

There is a compendium, as the *Elements* was, of Chinese knowledge. It originated between 200 BCE and 200 CE and was later amplified. The *Nine Chapters on the Mathematical Art* includes chapters on, among other topics, mensuration, properties of right angles, computation, taxation, and algebra. Its geometry included formulas for areas and volumes—some of them approximate, like the Egyptian and Babylonian—and approximations for π that went far beyond those of Archimedes.

2. Numeration

China had a strong central administration and extensive commerce. Both elements are spurs to the development of arithmetic. The result was place-value numeration and arithmetic of negative numbers and of fractions. In the latter two, China preceded Indian arithmetic, and may indeed have influenced it. (Recall that place-value numeration already existed in Mesopotamia.)

There were actually two systems of numeration. One was reminiscent of Egypt's, using decimal aggregates but not place value. It used markers for units and for 10 and its powers. Under it, if we use our numbers underlined as symbols for the corresponding markers, then we would write 54,321 as

5 10000 4 1000 3 100 2 10 1

and 2,012 as

2 1000 1 10 2.

We put the powers in descending order, but clearly decreasing order is unnecessary. Even order is irrelevant if we keep the coefficients with their powers. Therefore zero placeholders are unneeded. As always with reliance on aggregates, addition is easy and multiplication hard.

The other one amounted to a base-100 positional system. Recall again the disadvantage of the sexagesimal system: It required the Babylonians to symbolize the digits 1-59. The Chinese produced a way to render the digits 1-99 with 18 symbols, corresponding to 1, 2, ..., 9 and 10, 20, ..., 90. Then 54,321 would have its digits paired as 5/43/21 and would appear as

5 40 3 20 1;

similarly 2,012 would be

2 10 2.

Here the lack of a zero symbol is a problem, particularly for *trailing* zeroes (as in 2010). The zero the Chinese eventually adopted came *back* from India.

> Notice that both systems relate to the way many languages *name* numbers. In English, 11 and 12 have special names, but "thir/teen" through "ninety/nine" all reflect multiples of 10 plus some units. Spanish 11-15 are special, but 16 to 99 are ***dieci*/*séis*** to ***noventa y nueve*** (separate words). German has 11 and 12, but ***drei/zehn*** through ***neun/und/neunzig***. (French is—what can we say—Gallic.)
>
> Our own use of sort-of-base-100 names extends to 100 through 9999. We commonly call 4,321 "forty-three hundred twenty-one," although 4000 is rarely "forty hundred." The use stops at 9999; nobody calls 10,000 the equivalent of "one hectohundred."

Unlike the Egyptians, the Chinese dealt from early on with common (as opposed to unit) fractions. More interestingly, they dealt with decimal fractions more than a millennium before those came into use in India and the Islamic world.

3. Algebra

The algebra in *Nine Chapters on the Mathematical Art* is like what we know from the Egyptians and Babylonians, in that it presented a large set of specific problems and their solutions, generalization presumably left to the reader.

Chinese algebra may have been static through 1000 CE, but it certainly advanced in the next three centuries. By around 1270, Yang Hui (**Merzbach** says that we know almost nothing about his life) put the **binomial theorem** in a form recognizable to us. Recall that the theorem describes the expansion of powers of binomials: $(a + b)^n$ is a sum of terms; each term has a coefficient, a power of a, and a power of b; the powers of a decrease from n to 0, those of b rise from 0 to n; and the coefficients are given by **Pascal's Triangle**. Yang described the Triangle, perhaps known before him. (Named after the seventeenth century French scientist Blaise Pascal, it evidently should bear instead a Chinese name.)

The triangular array begins with 1, then expands as it descends, each position holding the sum of the entries to its upper left and upper right. Thus, in the table at right, the underlined 4 is the sum of the 3 and 1 above it. The red numbers are *row* numbers.

Row #1 tells us that

$(a + b)^1 = 1a^1b^0 + 1a^0b^1$.

Row #4 likewise says

$(a + b)^4 = 1a^4b^0 + 4a^3b^1 + 6a^2b^2 + 4a^1b^3 + 1a^0b^4$.

				1					0
			1		1				1
		1		2		1			2
	1		3		3		1		3
1		4		6		4		1	4

Yang also worked on a uniquely Chinese interest, magic squares A **magic square** is a square array of numbers, usually the integers from 1 to some integer square, in which all the rows, all the columns, and the diagonals have the same sum. To see the simplest possible example, try to fill a 3x3 array with the numbers 1-9 to render the correct sums. Exercise 1 asks you to attack the problem as a series of *deductions*. [Compare the recent Japanese puzzle import, *sudoku*.] The exercise demonstrates that the topic does not fall into the category of mere arithmetic. Yang wrote on considerably bigger squares.

The high point in Chinese mathematics generally and algebra in particular came in Zhu Shijie's book *Jade Mirror of the Four Origins* of 1303. The book treats problems describable by systems of simultaneous polynomial equations, where the systems reduce to single equations of high degree. Zhu then applies a method that now bears another European name, that of William Horner.

Horner's method has elements of false position. It works by reducing the constant in the equation.

Consider the simplest type of polynomial equation, which we could solve to extract a root:

$v^3 = 100$.

The substitution $w = v - 1$ will reduce the right side. Thus, $v = w + 1$ turns the equation (we put the binomial theorem immediately to work) into

$w^3 + 3w^2 + 3w + 1 \;=\; (w + 1)^3 \;=\; 100,$ or

$w^3 + 3w^2 + 3w \;=\; 99.$

Notice that the reduction on the right is 1^3.

The last equation is a more general type. It has the form

polynomial = (positive number),

where the polynomial has all positive coefficients and no constant. Write

$p(w) = w^3 + 3w^2 + 3w$.

Our example has

$p(w) = 99$.

[This is our first use of **function notation**. If it is unfamiliar to you, check dummies.com. No offense is intended. Strictly speaking, it is past the level this book is supposed to require.] Such an equation necessarily has a solution, because $p(0) = 0$ and p increases beyond 99. Here,

$p(3) = 63$ and $p(4) = 124$,

putting the solution between $w = 3$ and $w = 4$. [We are tiptoeing around the advanced principle called the "intermediate-value theorem."] If we now substitute $x = w - 3$, we write

$(x^3 + 3x^2[3] + 3x[3]^2 + 3^3) + 3(x^2 + 6x + 3^2) + 3(x + 3) = 99,$ or

$x^3 + 12x^2 + 48x \;=\; 99 - (3^3 + 3[3^2] + 3[3]) \;=\; 36.$

(Check that. Verify also that we would have ended up here if we had substituted $x = v - 4$ in the first place.) Notice that the reduction on the right is $p(3)$.

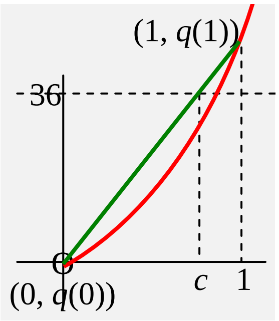
What is special about the last form is that the root is between 0 and 1. For that reason, the approximate solution is laid bare as

$x \approx$ (number)/(sum of coefficients) $= 36/(1 + 12 + 48)$.

That approximation is simply linear interpolation. Write

$q(x) = x^3 + 12x^2 + 48x$.

We have $q(0) = 0$ and $q(1) = 61$. The solution is therefore about $(36 – 0)/(61 – 0)$ of the way from $x = 0$ to $x = 1$.



Look at the interpolation graphically as well. On the right, we draw (red) the graph of $y = q(x)$ between $x = 0$ and $x = 1$. If the graph were a line (green), then the place $(c, 36)$ where it crosses the horizontal $y = 36$ would be given by similar triangles:

$36/c = q(1)/1$,

or $c = 36/q(1)$. The denominator $q(1)$ is the sum of the coefficients.

Lest we forget, the solution of the original equation is then given by

$100^{1/3} \approx 4 + 36/61$.

Exercises IV.B.3

1. Assume that the array at right is a magic square, *a*-*i* representing 1-9.

a	*b*	*c*
d	*e*	*f*
g	*h*	*i*

a) What is the sum of the numbers in each row?
b) Add the diagonals, the middle row, and the middle column to show that
$a + b + c + d + 4e + f + g + h + i = 60$.
c) Based on (b), show that *e* has to be 5.
d) Show that 9 cannot be at any of the corners. (Hint: If say $a = 9$, then the sums $b + c$ and $d + g$ must both be 6. There are not enough candidates left with sum 6.)
e) Set $b = 9$. Show that $a = 2$ and $c = 4$, or vice-versa.
f) Set *a* and *c* either way (2 or 4), then complete the array.
These show that the square is unique, except for symmetries (rotating it a multiple of 90° and/or flipping it.)

2. a) Write row #5 of “Pascal’s Triangle.”
b) Approximate $100^{1/5}$ by using “Horner’s method” on the equation
$x^5 – 100 = 0$.

4. Number Theory

In section III.B.5, we alluded to the **Diophantine linear system**

$x = 10y + 4$
$x = 21z + 5$

where x, y, and z are required to be integers. Qin Jiushao (13th century, whom **Merzbach** describes as an “unprincipled governor and minister”) dealt with such indeterminate systems in the *Mathematical Treatment in Nine Sections*. (He even treated systems of higher degree, and such geometric odds and ends as Heron’s formula.) We can read the above system as asking for an x that simultaneously has remainder 4 on division by 10 and 5 on division by 21. Then it falls under the next result.

Theorem 1. (The Chinese Remainder Theorem) Suppose the natural numbers m, n (and possibly others) are pairwise relatively prime (no two have a common divisor). Let r, s (and …) be corresponding remainders; that is, $0 \le r < m$, $0 \le s < n$, ... Then there are integers whose remainders are r under division by m, s under division by n, ...

In our example, 10 and 21 are relatively prime. (Check.) We know (Theorem 1 in section III.B.4) that we can find integers i and j such that

$$10i + 21j = 1.$$

Look at

$$x = 4(21j) + 5(10i).$$

For division by 10, we have

$$x = 4(1 - 10i) + 5(10i) = 10[5i - 4i] + 4.$$

Hence x has remainder 4, and

$$y = [5i - 4i]$$

solves the first equation. For division by 21, we have

$$x = 4(21j) + 5(1 - 21j) = 21[4j - 5j] + 5.$$

Hence x has remainder 5, and

$$z = [4j - 5j]$$

solves the second equation. (Verify this numerically in Exercise 1.)

In the spirit of Diophantus, we found one solution. In the spirit of Brahmagupta, let us characterize them all. There *are* others. After all, if we add a multiple of 10×21 to x, then the sum has the same two remainders. That turns out to be the *only* way to make other solutions.

Suppose that

$$s = 10t + 4$$
$$s = 21u + 5$$

supplies another solution. We will show that s has to be just x increased by some (possibly negative) multiple of 210, t has to be y increased by the same multiple of 21, and u has to be z increased by that same multiple of 10.

Subtract to write

$$s - x = (10t + 4) - (10y + 4) = 10(t - y) \quad \text{and}$$
$$s - x = (21u + 5) - (21z + 5) = 21(u - z).$$

The number $s - x$ is divisible by both 10 and 21. Therefore it is divisible by 10×21 (Exercise 2a). From $s - x = (10\times21)k$, we conclude that every solution s is just x plus some multiple $210k$. As for the rest of the solution,

$$(s - x)/10 = t - y \quad \text{forces} \quad t = y + 210k/10.$$

In words, t is y plus the like multiple $21k$. At the same time,

$$(s - x)/21 = u - z \quad \text{forces} \quad u = z + 210k/21;$$

that is, u is z plus the like multiple $10k$.

Extending the theorem to more divisors depends on just one more principle.

Check that 5, 21, 22, and 289 are pairwise relatively prime (Exercise 2c). Necessarily, the product of any of them is relatively prime to the product of any of the others. For example, 5(21) is relatively prime to 22, 289, and 22×289 (Exercise 2d). Therefore each is prime to the product of the other three. That means there are integer i’s and j’s with

$$21(22)289i_5 + 5j_5 = 1,$$
$$5(22)289i_{21} + 21j_{21} = 1,$$
$$5(21)289i_{22} + 22j_{22} = 1,$$
$$5(21)22i_{289} + 289j_{289} = 1.$$

From all those, we find that

$$r[21(22)289i_5] + s[5(22)289i_{21}] + t[5(21)289i_{22}] + u[5(21)22i_{289}]$$

has remainders r, s, t, u on division by 5, 21, 22, 289, respectively. (Compare Exercise3.)

Exercises IV.B.4

1. a) Find integers i and j such that $10i + 21j = 1$.
   b) Evaluate $x = 4(21j) + 5(10i)$, and verify that x has remainders 4 and 5 on division by 10 and 21, respectively.
   c) Find the smallest positive integer with those same remainder properties.
2. Show in (a)-(d) that:
   a) If v is divisible by a and b, **and a and b are relatively prime**, then v is divisible by ab. (One approach: prime factorization. A second: Write $v = ma = nb$ and work from there.)
   b) The conclusion in part (a) might fail if a and b are not relatively prime.
   c) The numbers 5, 21, 22, 289 are pairwise relatively prime. (Hint: prime factorization.)
   d) The product 5(21) is relatively prime to 22, 289, and 22×289.
3. a) Find i's and j's to make
   $21(22)i_5 + 5j_5 = 1$ (Hint: $3\times21(22)$ ends in 6, 1 more than a multiple of 5.)
   $5(22)i_{21} + 21j_{21} = 1$ (Hint: $5(22) = 5\times21 + 5$, so $17\times5(22) = 17\times5\times21 + 85$.)
   $5(21)i_{22} + 22j_{22} = 1$. (Hint: $5(21) = 5\times22 - 5$, so $13\times5(21) = 13\times5\times22 - 65$.)
   b) Evaluate
   $4[21(22)i_5] + 10[5(22)i_{21}] + 12[5(21)i_{22}]$
   and verify that it has remainders 4, 10, 12 on division by 5, 21, 22.

5. Astronomy

In China as in Greece, astronomy was important in driving mathematics. That is clear in the very title of the *Arithmetical Classic of the Gnomon and the Circular Paths of Heaven*. We normally think of a gnomon as the shadow-casting upright on the face of a sundial, a Chinese invention. However, a tall marked gnomon is usable to measure and chart the positions of stars. It can work directly to measure elevation, or work via timing to measure longitudinal location. Various courts built observatories with gnomons, quadrants (vertical quarter-circles), and other instruments that allowed charting of the skies.

They used observations to mark the months and the years. A month started when the thin crescent Moon first became visible in the dusk after New Moon. The year started with the winter solstice. When needed, they intercalated months to reconcile the lunar calendar with the seasons.

The extensive astronomical records of the Chinese led to two important, considerably later discoveries. Chinese records of comets reach back beyond the time of Jesus. They were part of the evidence that led Edmond Halley, in 1705, to conclude that a spectacular comet that appeared in 1682 had been appearing every 76 years or so for more than 2000 years. He inferred that the comet, which now bears his name, is a body in orbit around the Sun. Separately, the Chinese recorded a "guest star" in 1054 CE. It materialized suddenly, a star so brilliant that it was visible in daylight for months. In a few years, it faded so much as to disappear from the night. In the 1700's, Europeans discovered a faint "nebula" (Latin for cloud) where the guest star had appeared. It took 20th century photography to reveal that the nebula is a growing, glowing shell. The nebula is the expanding gaseous shell of a star that exploded. The shell's rate of expansion indicates that the explosion was the event of 1054. The "guest" was an inconceivably brilliant "supernova."

Section IV.C. Maya

Records are scarce for the Americans, the peoples whose lax immigration policies allowed hordes of undocumented Europeans to enter the lands on this side. The North Americans remained hunters until the encounter with the white men. The Andeans—our name "Incas" actually refers to the *rulers*, and they ruled only for the century before the Spanish came—had a civilization occupying much of western South America. Their empire covered nearly 2000 miles, from above Ecuador into Chile, between the mountains and the Pacific. They built cities and connected them with roads. They erected temples and citadels oriented to the directions of the solstices and equinoxes. Those sites' location, construction, and orientation evinced understanding of architecture, astronomy, even warfare. But with all this civil engineering knowledge, *they did not have a written language*. Only the remains of their structures give us evidence of what they achieved.

The one group of Americans for whose science we have documents is the Maya of southern Mexico and northern Central America. Their monumental architecture—temples, stadia, and the like—still survives. Our interest is a set of scroll-like books (**codices**, plural of **codex**) that reveal Mayan numeration and astronomy.

The Maya used a **vigesimal** system, almost. In such a base-20 system, you need digits for 0-19. Remarkably, they had a zero symbol centuries before the Indians did. The other digits were rendered with marks and 5-aggregates. Thus, 1, 2, ..., 18, 19 looked like

●, ●●, ..., ●●●///, ●●●●///.

The wrinkle came in the 20's place. There, the digit was limited to 0-17. The number

●●/// ●●●●/// = 17(20) + 19(1)

was followed by

● (zero symbol) (zero symbol) = 1(18×20) + 0(20) + 0(1).

The place values were therefore 1, 20, 18×20, 18×20^2, 18×20^3, ... (What is special about 18×20? See Exercise 1 for a related question.)

The Maya were careful trackers of the celestial wanderers, particularly of the Sun and Venus. Tracking the Sun led to accurate measurement of the year. Just like the Egyptians, though, they adopted a solar calendar of 365 days and accepted the drift of the seasons forward through the calendar. For Venus, they had extensive calculations of its positions, especially the heliacal risings.

[An astronomical highlight of 2012 was the transit of Venus. The planet crossed the face of the Sun, as seen from Earth, for the last time this century. It seems odd that the Maya would not have anticipated such events, but I have never seen any mention that they did. I would welcome the reader's guiding me to any source that speaks to the question.

I would equally welcome indication whether it was coincidence that a Venus transit—they are rare, occurring in pairs spanning eight years and separated by 105 or 121 years—should fall in the year the Maya predicted for the end of the world. That prediction made me worry I might not finish this book.]

Exercises IV.C.1

1. Place-value systems always use the sequence 1, b, b^2, ... of powers of a base b. But any sequence in increasing order will do.
   a) Argue why any natural number n of US cents can be assembled using the coins worth 1, 5, 10, 25, and 100 cents. In symbols,
   $n = 1a + 5b + 10c + 25d + 100e$,
   for some nonnegative $a \le 4$, $b \le 1$, $c \le 2$, $d \le 3$, e unlimited.
   b) Find an n that can be so expressed two different ways.
   c) Show that the expression *is* unique if we demand that the sum $a + b + c + d + e$ of the "digits" be as small as possible. That sum is the number of coins. Then check that making it minimal amounts to demanding $b + c < 3$.

Chapter V. The Road to Europe

Section V.A. The World of Islam

Muhammad lived 570-632 CE. Beginning in 610, he preached verses that were recorded as the Qur'an, which became the foundation of Islam. During his life, his followers took control of the Arabian Peninsula (modern-day Saudi Arabia and the countries to its south). After his death, they embarked on an extraordinary campaign of conquest. At its height, the resulting empire reached north to Turkey, east through Iraq, Persia, and India nearly to China, west through all of northern Africa to the Atlantic, north there to southern Spain, Sicily and other islands, and Greece.

Aside from extent, the campaign had another remarkable feature. The original conquerors were illiterate. They chose to absorb the knowledge and cultures of the conquered. Accordingly, they set scholars to translate books from the subjugated lands into Arabic (whose written form such scholars had created). They then extended the acquired knowledge into some of the most important scientific and mathematical discoveries of the fourteen centuries following Claudius Ptolemy.

[For the history of Arabic science, the outstanding book is Jim Al-Khalili's *The House of Wisdom*.

There is always difficulty in naming the people of this empire. The word "Arabs" is clearly a misnomer, since the empire encompassed Turks, Persians, Berbers, and many other peoples. "Muslims" is inaccurate, because the conquerors did not demand conversion to Islam. Not all in the empire knew Arabic, but it certainly became the language of the scholars, as Latin became in Europe. For that reason, we will follow **Merzbach** and **al-Khalili** and refer to "*Arabic* mathematics and science."]

1. The Translations

In 762, al-Mansur founded the city of Baghdad. Like Alexandria, it soon became an important commercial and cultural center. When paper was brought west from the China end of the empire, al-Mansur's successor established the world's first mills. Baghdad, again like Alexandria, became a capital of science.

Under Abdullah al-Ma'mun's rule, 809 to 833, Baghdad acquired the successor to the Museum, "The House of Wisdom." With the House began a golden age of translation that lasted two centuries. In fact, the campaign of translation was still going on at the western end, Spain, in the 1100's. The original scholars had translated eastern works, from Persia and India. Al-Ma'mun's interest in the *Elements* and *Almagest* led to work on the classics of Greece. From that, we have Ptolemy's famous work translated to Arabic by 845. (Ptolemy wrote two works called *Syntaxis*, Greek for "collection." The larger was called *E Megiste Syntaxis*, the Greater Collection. In Arabic translation, it became "The Greatest," *Almagest*. The Arabic title was not Ptolemy's idea.)

By early in the 900's, Diophantus and Aristarchus had been translated. By around 945, the long work on Euclid's *Elements* was complete. It is fair to say that much of the works of the Greeks would have been lost to us had they not been conveyed by Arabic scholars.

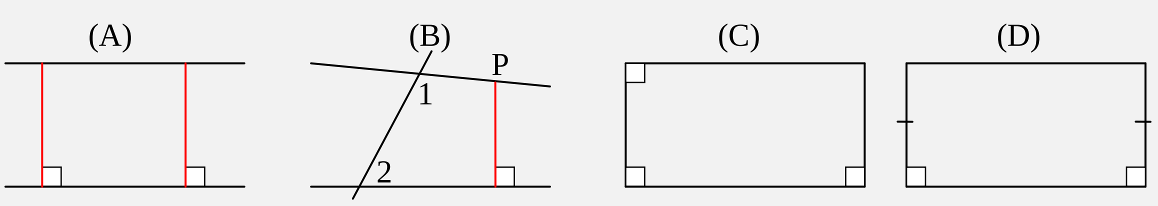
2. Arithmetic and Geometry

The greatest influences on Arabic mathematics were Mesopotamia and India. Consequently Arabic arithmetic and algebra developed much more than geometry. The arithmetic part was taken over completely from India. It was through the Muslim world that Indian numeration eventually got to Europe. But Ahmad al-Uqlidisi extended the numeration to decimal *fractions* around 950. Those became useful tools for Arabic astronomers, although their widespread use came five centuries later. (See **al-Khalili**, bottom of page 285, and compare with **Boyer**, page 268.)

In geometry, part of their legacy was carrying Indian trigonometry westward by around 1300. They completed our list of trigonometric functions, and even influenced our names for them. [See **Boyer** on the origin of "sine." But consult also **al-Khalili**, page 224.] It was not just a matter of outlook and names: Our forms of the laws of sines and cosines and of the double- and half-angle formulas were proved by Arabic mathematicians.

One area of Greek-influenced geometry was the attempt to prove the parallel postulate. Recall (section III.A.8b) that the question was to deduce the parallel postulate from the other Euclidean postulates. Arabic geometers pursued it by working on four propositions:

> a) *Parallel lines are necessarily equidistant.* In panel (A) of the figure below, we have two lines (black) that do not meet and the perpendiculars (red) to the lower line from two points on the upper. Prove that the perpendiculars are equally long.
>
> b) *If a transversal cuts two lines so as to form interior angles, on one side, that sum to less than a straight angle, then on that side the distance between the two lines reduces to zero.* In panel (B), angles 1 and 2 add up to less than a straight angle. Prove that the perpendicular (red) reaches zero length as point P recedes rightward.
>
> c) *If three angles in a quadrilateral are right angles, then so is the fourth angle.* That is the picture in panel (C), where one must prove that the remaining angle is a right angle.
>
> d) *If in a quadrilateral, two opposite sides are congruent and perpendicular to a third side, then the quadrilateral is a rectangle.* In panel (D), the vertical sides are congruent, and they are perpendicular to the base. It suffices to prove that the top is congruent to the base, or that it is perpendicular to one of the verticals.



[**Boyer** ascribes (d) to Omar Khayyam and (c) and (a) to ibn al-Haytham; see both later.]

The parallel postulate implies each of these. For example, Exercise III.A.8b:2 asked for proof that if you assume the parallel postulate, then the equidistance (proposition (a)) follows. The exercise below addresses the other three statements. It is harder to prove that each of them implies the postulate, but it can be done. They are all, therefore, equivalent to the postulate. However, none of them actually follows from the other postulates. The geometers ended up in the same trap as the Greeks, relying for proof on other assumptions equivalent to the postulate.

Exercises V.A.2

1. Assume the parallel postulate.
   a) Prove propositions (b) and (c).
   b) Prove that in panel (D) of this section's figure, the top side is congruent to the base and perpendicular to the two vertical sides.

3. Algebra

Muhammad al-Khwarizmi lived about 800-850. He was a geographer and astronomer of such renown as to be invited to head the House of Wisdom. (Think of Eratosthenes at the Library a thousand years before.) With that description, you might figure that he contributed to geometry, as he did. He also wrote a book on Indian numeration that eventually transmitted "Arabic" numbers (but, oddly, not negative numbers) to Europe. (See Boyer.) However, he is best known for creating what came down to us as algebra.

His most important book is called *al-Jebr*. Part of it followed Egyptian and Babylonian precedent in dealing with geometric questions: areas, land distributions, and the like. It also separated forms of equations into cases. But it introduced, for the first time, statements of general methods that led to solutions for the cases. The methods are what we now call *solution algorithms*. Algebra as *theory of equations*—as the study of methods for the production and solution of equations—began with this book.

[Words like "alcohol," "almanac," and "Alberto" evince their Arabic origins by their first two letters. Doubtless you have already seen that "algebra" comes from the book, and "algorithm" honors the scholar.]

To illustrate "cases" and "general methods," consider how we deal with quadratics. What we call the "quadratic formula" addresses the general form

$$ax^2 + bx + c = 0,$$

in which a, b, c can be any real numbers. Al-Khwarizmi would not have dealt with that form (Exercise 3). He would have turned

$$x^2 + 12x - 70 = 0$$

into $x^2 + 12x = 70.$

The long name of *al-Jebr* included words for "comparing" or "balancing." [The long name is *al-Kitab al-Mukhtasar fi Hisab al-Jebr wal-Muqabala*. "That's easy for you to say, you might be thinking," adds **al-Khalili**, page 110.] The latter turns up in the idea of adding the same quantity to both sides. We did just that to produce the last equation, which requires no subtraction or negative coefficients. That equation is the case of the quadratic in which the constant term is isolated. Two other cases would be

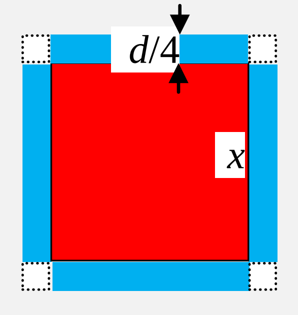
$x^2 + e = dx$ (linear term isolated)

and $x^2 = dx + e$ (quadratic term isolated).

(Al-Khwarizmi also considered the three cases in which a term is missing. We can skip those. If the quadratic term is not there, then it is not a quadratic equation; if the linear term is missing, the question is simply to find a square root; and if the constant term is out—and zero is not considered a solution—then division reduces the equation to a linear one.)

For the first case, al-Khwarizmi laid out a method that was already known to the Babylonians. It is worth our attention because we still use it.

> Given the form
>
> $$x^2 + dx = e,$$
>
> draw a square (red in the figure below right) of side x. Extend each side by 1/4 of the x-coefficient, adding $d/4$ to each length. Then draw the horizontals and verticals to enclose four rectangles (blue) sized $d/4$ by x. At this point, the colored area is
>
> $$x^2 + 4(d/4)x = x^2 + dx,$$
>
> which we know to be e. If we enclose the missing corner pieces (by the dashed outlines), we add four areas $d^2/16$ and produce a new big square. The big square has sides $x + 2(d/4)$ long. That means
>
> $$(x + d/2)^2 = e + 4(d^2/16).$$
>
> We take the root of interest, the positive one, to write
>
> $$x + d/2 = \sqrt{[e + 4(d^2/16)]}.$$
>
> The value of x is clear.



The method is named for what it does, **completing the square**. Recall that we teach it in school algebra because it proves to be useful here and there, but its first application is in the derivation of the quadratic formula.

A worthy successor to al-Khwarizmi was Umar al-Khayyami, the Persian poet Omar Khayyam (1050-1123). His book *Algebra* ventured beyond *al-Jebr* into the first systematic treatment of solution of cubic equations. He solved them in terms of intersections of conic sections, more reminiscent of the methods of Archimedes and Apollonius than of al-Khwarizmi. (See Exercise 5 for illustration.) On the other hand, he did organize them into cases. That is considerable work, because there are many cases; see Exercise 6.

Exercises V.A.3

1. (**Boyer**) a) Compare, in their effect on learning, the Arabic conquests with the earlier ones of Greece and Rome.
   b) Name parts of Greek mathematics that would have been lost to us, except for their transmission within the Muslim world.
   c) In what ways would the Crusades have helped, and in what ways hindered, the transmission of Islamic math to Christian Europe?
2. How did the Greek approach to geometry differ from Al-Khwarizmi's approach to algebra?
3. (After **Boyer**) Why did al-Khwarizmi's algebra not treat quadratics of the form
   $ax^2 + bx + c = 0$
   (equivalently, $x^2 + dx + e = 0$)?
4. Solve $x^2 + 12x = 70$ in the (pictorial) style of al-Khwarizmi. Compare with the solution given by our quadratic formula.
5. Put the conic-section approach into modern terms:
   a) Sketch the parabola and hyperbola given by
   $y = x^2 + 2x + 10$ and $y = 20/x$.
   b) Argue why they have exactly one intersection.
   c) Argue why the x-value at the intersection is the solution to the cubic
   $x^3 + 2x^2 + 10x = 20$.
   d) Use the graphs to approximate the solution to about 0.1.
6. We implied that our form of quadratic equation,
   $ax^2 + bx + c = 0$,
   rearranges into six possible cases of an equation with positive coefficients and positive roots. They come from six possibilities: any (single) one of a, b, and c being zero and the other two having opposite signs; or one of them having sign opposite that of the other two. How many such cases can the cubic
   $ax^3 + bx^2 + cx + d = 0$
   turn into?

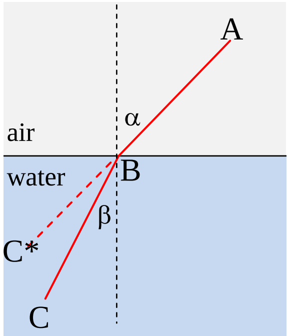
4. The Sciences

We saw in Section III.C that in the four centuries from Aristarchus to Ptolemy, Greek astronomy achieved its greatest triumphs. Copernicus came 1400 years after Ptolemy, and may fairly be said to mark the reawakening of astronomy, and science in general, in Europe. In between, the heights of science were reached in the Islamic world. The discoveries in chemistry, medicine plus physiology, physics, and philosophy represented the limits of human knowledge for at least five centuries.

We will look at two examples that are by no means the most important discoveries, but that are amenable to study by means of our geometry.

a) optics

In optics, Arabic discoveries included explanations of vision and the action of lenses. These appeared in the *Book of Optics* of Abu Ali al-Hassan ibn al-Haytham (965-1039). It was the most important work in the subject since Ptolemy's book more than eight centuries earlier. Understanding of lenses led to the invention of telescopes. Those finally reached Europe, via the Dutch, in the 1500's.



The explanation of lens action relied on a discovery that preceded ibn al-Haytham, the **law of refraction**. In the figure at left, we have a ray of light (red) from A, entering a body of water at B, continuing to C. It was known that such a ray bends (**refracts**) away from the surface. Science describes the phenomenon by reference to the **normal**, the perpendicular (dashed) to the surface. The ray bends toward the normal, so that the **angle of refraction** β is smaller than the **angle of incidence** α. The process works as well in the opposite direction. A ray from C surfacing at B refracts away from the normal, to A. At A, our brains would interpret the ray as coming in a straight line from C*. That is why a fish or stone in the water appears to us to be closer to the surface than it actually is.

Ptolemy thought that β would be proportional to α: β/α = constant. From considerations of speed of propagation (Exercise 1), we now know that the relation is actually

$$\sin\beta/\sin\alpha = \text{constant}.$$

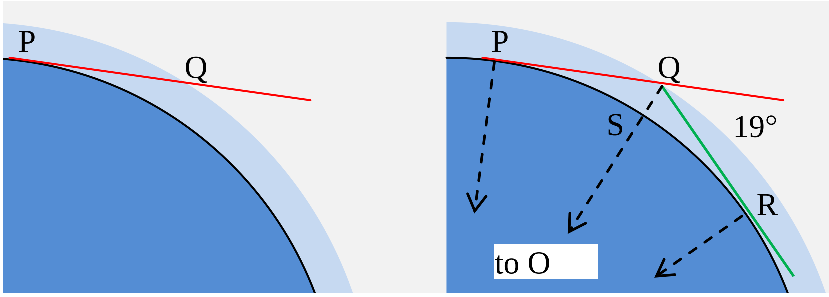
(Notice that Ptolemy was approximately right for small angles. For them, $\beta/\alpha \approx \sin\beta/\sin\alpha$.) The "constant" is independent of the two angles, but it depends on the two *media*. It is the ratio of speed of light in water to speed of light in air. If you substitute a body of glass for the water—or for the air—then the speed in the corresponding medium changes, along with the constant ratio. What is more, the constant depends on the *color* of the light. Refraction increases (smaller constant) as you move from the red end of the spectrum to the violet.

The relationship is called "Snell's Law," after a European. (What else? He was Dutch, living either side of 1600.) But the first one to describe it was al-Ala' ibn Sahl, 600 years before. The description took a delightful form; see **al-Khalili**, page 163. Ibn Sahl described refraction *in terms of half-chords*. (In the figure above, which assumes that AB and BC are equally long, the half-chords would be proportional to the perpendiculars from A and C to the normal, just like the sines of α and β.) In other words, he used the Indian version of trigonometry to render the law of refraction.

b) astronomy

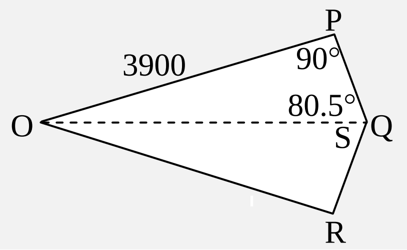
In 828, al-Mansur commissioned the establishment of an observatory at Baghdad. It was originally meant to extend the observations of Ptolemy, and it produced important star charts. Part of its legacy is still in the sky: Many stars, Betelgeuse perhaps most famous, have names from Arabic. From Arabic astronomy, there is one discovery we will chase for its geometric value. It is ibn Mu'adh's calculation of the height of the atmosphere.

In the left half of the figure below, we see Earth (blue) surmounted by a layer of air (lighter blue). For an observer standing at P, the line of sight to the horizon is roughly the tangent (red) at P to the planet. The tangent reaches the (assumed) top of the atmosphere at Q and continues into space. When the Sun is above PQ, there is daylight at P. When Sun's leading edge reaches PQ, sunset at P begins. [Ignore the "sunset illusion." The atmosphere's thickness increases as you go down. Hence it refracts the light of the setting Sun—and moon, stars, any sky object near the horizon—toward the normal; in a word, downward. Our brains, always interpreting the light as coming in a straight line, “see” the Sun about ½° higher than it actually is. Since the Sun happens to be ½° across, when the Sun appears to be sitting on the horizon, it is actually sitting just *below* the horizon.]



In the right half (after **al-Khalili**, page 165), we add the other tangent (green) from Q, meeting Earth at R, and the lines (dashed) from P, Q, R toward the center O of Earth (out of the figure). OQ crosses the surface at S. The astronomer ibn Mu’adh reasoned that while the Sun is above QR, it illuminates some of the atmosphere through QS. That makes the western sky display twilight to the observer at P. When the Sun gets below (the extension of) QR, it no longer illuminates any of QS, and twilight ends for P. He further estimated that twilight ends when the Sun reaches 19° below the horizon. In other words, the angle between PQ and QR is 19°.

[The moment when the Sun reaches an **angle of depression** of 18° is called the **end of astronomical twilight**. It represents a modern convention that serious observing of the night sky can start then. It does not happen at a fixed time after sunset. Our (temperate zone) experience is that summer twilight lasts longer than winter. The reason is that the summer sun approaches the horizon at a slant; it descends slowly below the horizon. The winter sun approaches the horizon at almost a right angle; it descends quickly. For his estimate, then, Ibn Mu’adh needed to consider the rate of descent.]



With those numbers, Ibn Mu’adh ended up with the figure at right. The quadrilateral has the technical name “kite”: It has two pairs of congruent adjacent sides. OP and OR are congruent because they are Earth radii, for which astronomers (including al-Biruni) had calculated around 3900 miles. QP and QR are congruent because they are tangents from a common point. They also have to be perpendicular to the radii, and OQ bisects the angles at O and Q. Since angle PQR is

$$180° - 19° = 161°,$$

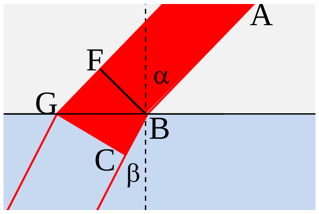
triangle OQP has angle OQP = 80.5° opposite 3900. Therefore

$$3900/OQ = \sin 80.5°.$$

We calculate OQ ≈ 3954 miles. That leaves SQ ≈ 54 miles for the height of the atmosphere.

There is no real end to the atmosphere, but beyond that altitude (just past the conventional end of the stratosphere) there is less than a millionth of the air. Go back to Section III.C to compare this wonderful combination of observation and geometric reasoning to those of Aristarchus and Eratosthenes.

Exercises V.A.4



1. In the figure at right, a red beam of light approaches the water at angle of incidence α, then continues through the water at angle of refraction β. Its edges AB and GF are parallel, and are perpendicular to the “wavefront” BF. In the time t after the front reaches B, the upper edge of the beam travels FG through the air to the water at speed V; the lower edge travels BC through the water at speed v; and GC becomes the new front, perpendicular to BC. Prove that

$$\sin \beta / \sin \alpha = v/V.$$

Section V.B. Medieval Europeans

A great deal of interaction between Christian Europe and the Muslim dominions consisted of war. That is true even if you discount the Crusades, 1095 to 1291. The Moors conquered Andalusia, roughly the southern quarter of Spain, in 711. They held most of it until the natural stronghold of Toledo was recaptured by Alfonso VI in 1085, and were not driven completely out of Spain until 1492. By then, they were also out of Sicily (but not Greece, which Turkey ruled until the 1800s.)

1. The Translations

During their Andalusian hegemony, the Moors established at Córdoba a center of science in the style of Baghdad. Oddly, the forces driving this science were medicine and philosophy, as opposed to the astronomy and mathematics of Baghdad. (See **al-Khalili**, chapter 13. **Struik** says that the large-scale, unirrigated agriculture of the West never provided a stimulus to astronomy.) Its golden age was 929-1031. Early on, interaction with Europeans began to include scholars. The Christians who journeyed to Toledo and Córdoba—and to smaller extents, to Venice (which had much trade with the east) and Sicily—recognized the importance of the Arabic texts. They started to translate them into Latin. By about 1150, there existed Latin versions of the *Elements*, *Almagest*, and most interestingly, *al-Jebr*. Its translation made an impression on the first great Western mathematician.

2. Leonardo of Pisa

[I ask the reader to indulge my preference for his given name. Leonardo of Pisa is generally called “Fibonacci.” That is just a nickname meaning “son of Bonaccio.” The last was, in turn, not his father’s name, but a nickname meaning “good guy.” Calling him Leonardo of Pisa distinguishes him from the later genius Leonardo the Florentine, who of course is almost always called by his place-name, da Vinci.

I would have far preferred to call Abu Abdullah Muhammad ibn Musa al-Khwarizmi—Muhammad from Khwarizm, father of Abdullah, son of Moses—by his given name. That would have created the obvious problem that the name is best reserved for The Prophet. The greater problem is that al-Khwarizmi is terribly important, rates therefore widespread mention, and is universally identified by that place-name. Prof. al-Khalili was kind enough to answer my query about naming. He pointed out that Leonardo referred to al-Khwarizmi as “Maumeht,” the Latin version of the name, but added a guide that I will mostly follow: “I just use the name he is best known by.”]

a) numeration

Leonardo lived roughly 1180-1240. His greatest work was the *Liber Abaci*, 1202. The title might better be rendered “Book of Calculation” than “Book of the Abacus,” because one of its themes was the advantage for calculation of decimal numeration. Exposed in his travels to the Latin translation of al-Khwarizmi’s book on numeration, Leonardo tried to popularize the use of Indian numbers (which had been introduced to Europe before.) Perhaps he saw translations of Brahmagupta as well, because he dealt with the arithmetic of negative numbers. Those would have been an important tool in commerce, since the beginnings of banking and credit (and what you might call currency exchange) trace back to Leonardo’s time.

b) algebra

Leonardo's most famous contribution is the **Fibonacci numbers**. They form a sequence that begins with 1, 1, then sets each subsequent term to the sum of the previous two. Thus, the sequence looks like

$$1, 1, 2, 3, 5, 8, 13, \ldots.$$

(You can start with any two numbers. As long as the **recursion**—the definition of new terms by reference to the ones already there—stays

$$f_{n+2} = f_{n+1} + f_n,$$

the result is called **a** (**generalized**) **Fibonacci sequence**.)

Leonardo used the sequence in answer to a specific question. Suppose you have a pair of rabbits, and they mature to mate at the end of month #1. At the end of every subsequent month, they produce a pair of offspring. The descendants then behave like the original: a new pair of young at the end of every month, beginning with the second month of life, forever. How many pairs are there at the start of each month? The numbers born to that schoolbook exercise turned out to have a remarkable number of algebraic properties (Exercises 1-2), plus a huge number of appearances in nature. Except for the two exercises, we will not pursue them. The linked article has a wealth of references where you can chase.

More relevant to our interest is that *Liber Abaci* introduced to Christian Europe the algebra of the worlds of Islam and India. It teaches the methods of al-Khwarizmi, expanding on them to allow negative solutions. Recall that the problems of al-Khwarizmi had everyday settings, like questions about land areas and inheritances; he did not allow answers of zero or less. Leonardo was pitching commercial applications, where negative numbers could be readily interpreted as debts. The *Liber* also teaches and expands on Brahmagupta's methods in indeterminate equations.

It even expands on Omar Khayyam's work on cubic equations. Look at

$$x^3 + 2x^2 + 10x = 20,$$

which is sometimes called "Fibonacci's cubic." Omar had expressed the opinion that some cubic and higher-degree equations could not be solved by combinations of roots. Leonardo drew the conclusions of Exercise 4, plus the next step, that even such a combination as

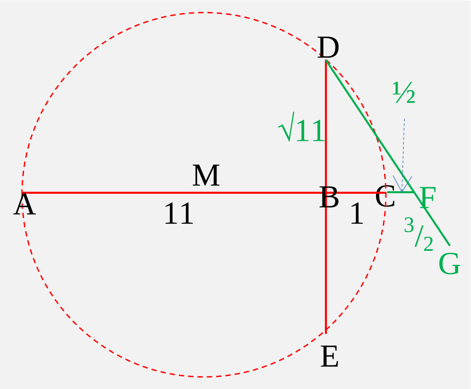
$$\sqrt{(m/n + \sqrt{[k/l\,]})}$$

would not solve this equation. Such combinations represent lengths we can construct with straightedge and compass. That constructibility is in contrast to the solution in Exercise V.A.3:5; you cannot construct parabolas and hyperbolas with just straightedge and compass.

Sometimes you can construct a solution without solving its equation. We illustrate with

$$x^4 = 3x^2 + 11.$$

> Draw AB of length 11 and extend it 1 past B to C. Find the midpoint M of AC, and draw the circle of radius AM centered at M. Erect the perpendicular to AC at B, meeting the circle at D and E. The segments and circle are red in the figure. By one Euclidean theorem, the products of the pieces of intersecting chords are equal. Thus,
>
> $$DB \times BE = AB \times BC = 11.$$
>
> By a second theorem, the diameter AC bisects any chord perpendicular to it. Thus,
>
> $$DB = BE.$$
>
> That gives us
>
> $$DB^2 = 11.$$
>
> (This is a standard construction; it produces the square root of any given length. Compare section III.A.3b.)



> Now extend BC to make BF = 3/2, half the *x*-coefficient. Draw DF and extend it past F, to G, by the same 3/2. All those additions are in green. In right triangle DFB, we have
>
> $DF^2 = (3/2)^2 + (\sqrt{11})^2$.
>
> That yields
>
> $(DG - 3/2)^2 = 9/4 + 11,$ or
>
> $DG^2 - 3DG + 9/4 = 9/4 + 11,$ and
>
> $DG^2 = 3DG + 11.$
>
> If we now construct the square-root length GH, so that $GH^2 = DG$, then we have
>
> $(GH^2)^2 = 3GH^2 + 11.$
>
> We have constructed the solution GH without solving algebraically. (Verify in the figure that GH has to be $\sqrt{(3/2 + \sqrt{[11 + 9/4]})}$. Compare that with the answer to Exercise 6.)

Observe that Omar and Leonardo's thought here is unlike earlier algebraic thought. It does not relate to practical problems; cubic equations rarely do. It does not specifically try to create solution algorithms or lines of attack. Instead, it focuses on the *nature* of unspecified solutions. Algebra would move increasingly in that direction in the coming centuries.

Exercises V.B.2

1. Show that any two consecutive Fibonacci numbers are relatively prime.
   (Hint: Any number that divides f_{50} and f_{51}, say, must divide all the previous ones also.)
2. The ratio f_{n+1}/f_n of consecutive Fibonacci numbers has a limit *L*. (Without the calculus language: There is a number *L* with the property that for all large values of *n*,
   $f_{n+1}/f_n \approx L$,
   and the approximation gets closer to exact as *n* increases indefinitely.)
   a) Figure out *L*. (Hint: Write the recursion as $f_{n+1} = f_n + f_{n-1}$.)
   b) Where did we meet that number before?
3. Consider an organism that is born one day, produces one offspring the next, produces another offspring the third, then dies. (This is more realistic than Leonardo's immortal rabbits. Everybody knows that dealing with multiple young is bound to kill you.) The offspring do the same.
   a) How many creatures are born on day *n*?
   b) How many are alive at the end of day *n*?
4. For Fibonacci's cubic equation
   $x^3 + 2x^2 + 10x = 20$:
   a) Show that if the reduced fraction m/n, $n > 0$, satisfies it, then $n = 1$.
   (Hint: If *n* is relatively prime to *m*, then it is relatively prime to all powers of *m*.)
   b) Show that no integer satisfies it.
   (Hint: Eliminate negative integers, then 2 and above, then 0 and 1.)
   c) Use (a) and (b) to show that the equation has no rational solution.
   d) Show that the simple irrational $\sqrt{(m/n)}$ cannot solve the equation either.
   e) Show that even $m/n + \sqrt{[k/l]}$ will not work.

5. (Calculus)
   a) Prove that Fibonacci's cubic must have a real solution.
   b) Prove that it has exactly one solution.
   c) Prove that the solution is between $x = 1$ and $x = 2$.
   d) Calculate to approximate the solution to within 0.01. ("Calculate" means you may use the arithmetic functions, but not the algebraic or graphics power, of a calculator.)
6. Use the quadratic formula to find the lone positive solution of
   $x^4 = 3x^2 + 11$.

3. Oresme

The most accomplished algebraist of the century after Leonardo was the cleric Nicole Oresme [aw-REM; French is wasteful with letters], around 1320-1382.

a) fractional powers

One of his original ideas was fractional powers. We can describe these as intermediate proportionals. Thus, half powers fit halfway between whole-number powers by

$$a^{3/2}/a \;=\; a^2/a^{3/2} \;=\; a^{5/2}/a^2 \;=\; a^3/a^{5/2}.$$

Notice that the first proportion implies

$$(a^{3/2})^2 = a^3.$$

That meshes with our definition

$$a^{3/2} = \sqrt{}(a^3),$$

which we can match up with $(\sqrt{}a)^3$. Similarly,

$$a^{5/4}/a \;=\; a^{6/4}/a^{5/4} \;=\; a^{7/4}/a^{6/4} \;=\; a^2/a^{7/4}$$

defines fractional powers of denominator 4 (with such caveats as Exercise 1).

(Jordan of Nemore, who overlapped with Leonardo, was first to use letters to represent quantities. That step is important, because it allows concise statements of algebraic principles. Consider how

$$(a + b)^2 \;=\; a^2 + 2ab + b^2$$

renders a general law much more simply than the corresponding words would.)

b) series

Oresme made contributions to the study of "infinite sums." It was known that such things might or might not make sense. Clearly we cannot interpret 1 + 1 + 1 + 1 +… as a number. It represents something bigger than any fixed number; we would call it infinity. On the other hand, Archimedes had "summed" an infinite series in the quadrature of the parabola (Section III.A.6(iv)), and there are good reasons to write

$$1 + 1/2 + 1/4 + 1/8 + \ldots = 2.$$

> For one thing, that "sum" is not bigger than 2. That is, it never reaches 2, because its **partial sums**
>
> $$1,\ 1{+}1/2,\ 1{+}3/4,\ 1{+}7/8,\ \ldots$$
>
> are all less than 2. At the same time it is not less than 2, because its partial sums eventually exceed any number that *is* less than 2 (Exercise 2). For another, write
>
> $$s = 1 + 1/2 + 1/4 + 1/8 + \ldots.$$
>
> Then
>
> $$2s \;=\; 2 + 1 + 1/2 + 1/4 + 1/8 + \ldots \;=\; 2 + s.$$
>
> That forces $s = 2$. (Dividing by the common ratio—multiplying by its reciprocal—is the standard way to evaluate the sum of a geometric series. See Exercise 3 for a variation.)

On the third hand, manipulating series as though they were actual (finite) sums leads to paradoxes. Write

$$t = 1 - 1 + 1 - 1 + \dots.$$

From

$$t \;=\; 1 - (1 - 1 + 1 - 1 + \dots) \;=\; 1 - t,$$

we conclude $t = 1/2$. If instead we interchange terms 2 and 3, terms 4 and 5, and so on, to write

$$t \;=\; 1 + 1 - 1 - 1 + 1 + 1 - \dots \;=\; 2 - 2 + 2 - 2 + \dots \;=\; 2t,$$

then we conclude $t = 0$. We will not settle on a meaning for infinite series for another 500 years.

Between the extremes, we have the subtle **harmonic series**

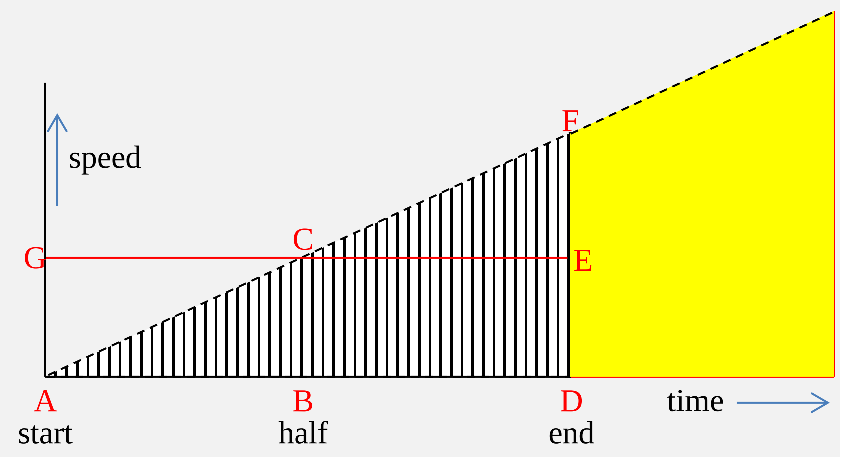
$$1 + 1/2 + 1/3 + 1/4 + \dots.$$

Does that infinity of vanishingly small terms add up to a number, like the $1/2^n$ and $(1/4)^n$ progressions? Oresme gave an incontrovertible argument that it does not.

> Look at the terms from just after a power of 1/2 until the next power. Of the two terms 1/3 and 1/4, the latter is smaller, so that
>
> $$1/3 + 1/4 > 2/4.$$
>
> Of the four terms 1/5, 1/6, 1/7, 1/8, the last is least, making
>
> $$1/5 + 1/6 + 1/7 + 1/8 \;>\; 4/8.$$
>
> Similarly
>
> $$1/9 + 1/10 + 1/11 + 1/12 + 1/13 + 1/14 + 1/15 + 1/16 \;>\; 8/16, \dots.$$
>
> Therefore
>
> $$1 + 1/2 + 1/3 + 1/4 + \dots \;>\; 1 + 1/2 + 1/2 + 1/2 + \dots,$$
>
> and the last is clearly infinite. (See Exercise 4.)

c) mechanics

Finally, Oresme investigated a question in the study of motion. (Before him, Jordan had already made a great contribution to mechanics [**Boyer**], and others were studying defects in the mechanics of Aristotle.) The subject was the distance covered by an object starting from rest and accelerating uniformly, meaning gaining speed at a constant rate.



Oresme, three centuries before Descartes, made the equivalent of a graph of the speed vs. time. At equally-spaced points (marking “longitudes”) along the horizontal time line, he raised vertical segments (shown black in the figure) whose heights (“latitudes”) corresponded to the current speed. Since the verticals are uniformly spaced in time, the heights are also proportional to the distances covered during the intervening short, equal time spans. Viewing the infinity of possible segments as constituting the region under the sloping dashed line, Oresme in effect saw the distance as an area. [You could make the same argument for any variation in the speed, not just uniform increase. Oddly, none of the histories suggests that Oresme made the generalization.]

This geometric interpretation yields a number of conclusions. Add the red letters and horizontal midline GCE to the figure. Triangle CEF is congruent to triangle CGA. Therefore the whole area under AF equals the area under the red line. That says the distance covered is the same as would be covered at a constant speed equal to the speed BC halfway through the interval (see Merton rule). Furthermore, rectangle BDEC has twice the area of triangle ABC, so that the area under CF is three times the area under AC. Similarly, continuing the graph to the next subinterval of length BD encloses an area (yellow in the figure) five times the area under AC. The distances for consecutive subintervals are therefore in the ratios 1: 3: 5: …. That is itself an important discovery, but it also means that the total distances

$$1, \qquad 1 + 3 = 4, \qquad 1 + 3 + 5 = 9, \qquad 1 + 3 + 5 + 7 = 16, \ldots$$

are proportional to the squares of the time. Those became Galileo's discoveries, around 1600, from his experiments with falling bodies.

Exercises V.B.3

1. Is this true for all real numbers r:

   $\sqrt{}(r^3)$ (which is the definition of $r^{3/2}$) is the same as $\sqrt[4]{}(r^6)$ (definition of $r^{6/4}$)?

2. What is the first power n such that

   $1 + 1/2 + 1/4 + \ldots + 1/2^n$

   exceeds 1.999 999 999?

3. Let

   $u = 1 + 1/4 + (1/4)^2 + (1/4)^3 + \ldots.$

   Subtract 1 from both sides, then factor 1/4, to evaluate u.

4. For what values of n is

   $1 + 1/2 + 1/3 + 1/4 + \ldots 1/n > 1{,}000{,}000{,}000$?

   (Hint: Estimate from either Oresme's argument or from the integral of $1/x$.)

Chapter VI. Renaissance Europeans

In the 1400's, Europe began to assume the leading role in the sciences and mathematics. **Boyer** suggests an interesting interpretation: You could date the beginning of the new age to the 1436 birth of Regiomontanus. At that point, Europe was recovering from the disaster of the Plague, commerce was fueling relative prosperity, and printing was just a few years away. National states were forming, and by century's end would embark on the Age of Exploration. Having absorbed Arabic scientific knowledge, Europe became the center of discovery just as Arabic scientific inquiry was ending.

In mathematics, the Arabic influence meant that algebra was of far greater interest than geometry.

Section VI.A. Geometry

The principal geometric advances came in trigonometry, but two other pursuits came under study. Those two were related to problems of capturing three-dimensional features on flat media, like paper.

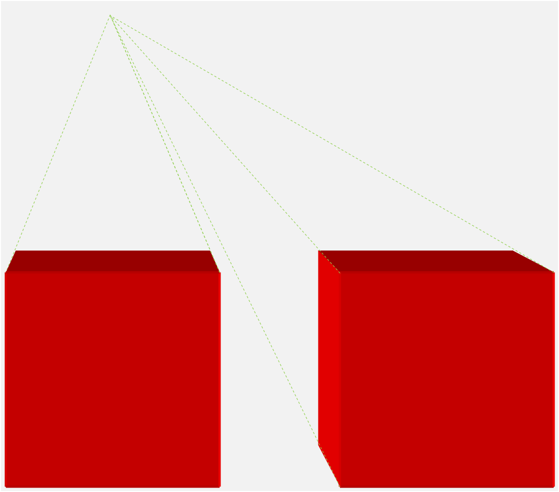
1. Trigonometry

Regiomontanus trained in astronomy, which of course immersed him in trigonometry. (He took a Latin name meaning "King's Mountain," because it translated the German name *Königsberg*, where he was born Johann Müller.) His book *De Triangulis Omnimodis* ("Of Triangles of All Kinds") included many theorems and examples on the solution of triangles. He used properties of right triangles, moving toward our approach. (Compare Exercise 3.)

Regiomontanus saw the potential of the printing press to disseminate mathematical and scientific knowledge. Unfortunately he died in 1476 (aged 40), before the appearance of even *De Triangulis*. Its publication (1533) established trigonometry as a separate area of inquiry (and not just a tool of astronomy). It appeared in time to come to the attention of Copernicus. He also wrote a trigonometry book, obviously influenced by Regiomontanus. Copernicus's student Rheticus created one of his own, *Opus Palatinum de Triangulis* ("Work on the Science of Triangles," the name "trigonometry" being in the future). It had elements that appeared for the first time: tables of all six of our trig functions, and their expression in terms of sides in a right triangle.

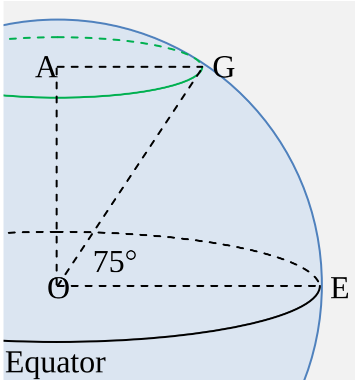
2. Perspective and Cartography

The development of **perspective** was driven by art. It is frequently said that the most striking difference between medieval and Renaissance art is the rendering of perspective. To take one rudimentary part of it, consider the "cubes" drawn at right. In this picture, the parallel left and right edges of the closer cube's square top appear to converge (along the dashed green lines) toward a "vanishing point," in the distance behind the cube. For the farther cube, the edges also seem to converge. But from our point of view—from our perspective—they converge not behind that cube, but leftward toward the same vanishing point as for the first cube. The three-dimensional effect is heightened by the coloration of the solids, suggesting that the light is coming from the left foreground. Numerous artists wrote on the mathematics involved, but its strictly mathematical treatment (**projective geometry**) did not come until the late 1600's.



The other pursuit was cartography. Here the most important name is Mercator (1512-1574. The name is Latin for "merchant." It translates his Flemish surname; he was born Gerard de Cremer.)

In mapmaking, the problem is to image a solid Earth on a flat sheet. One approach is to define lines of latitude and longitude on Earth's surface, plot a grid of lines on the paper, and represent land features on the paper according to their coordinates. It is irreproachable at the scale of tens of miles, satisfactory at a few hundred miles. At continental or global scale, however, a grid of equally-spaced lines overstates east-west distance. That distortion increases toward the poles. The reason is that Earth's meridians of longitude converge as you go north or south away from the Equator; indeed, they all meet at the poles. Look at this Mercator-derived map. Its longitude lines (vertical, at multiples of 15°) are equally spaced. Therefore it shows Greenland, at its widest, slightly less wide than the northern USA. In reality, that part of Greenland is about 760 miles across, the US 2850 miles.



> In the figure, we have a spherical Earth (blue) girdled by the Equator and by the circle ("parallel," green) of latitude 75°. The latitude is the angle GOE between the line OG, from Earth's center O to G on the parallel, and the plane of the Equator. The Equator (black band) has circumference $2\pi(\mathrm{OE})$. If A is the center of the parallel, then OAG is a right triangle. We gauge that the parallel has radius
>
> $\mathrm{AG} \;=\; \mathrm{OG}\cos 75° \;=\; \mathrm{OE}\cos 75°,$
>
> giving it circumference (2π OE cos 75°). Therefore the map, displaying the 75th parallel (first horizontal line below the top of the map) with the same length as the Equator, exaggerates the parallel's length by a factor of $1/\cos 75° \approx 3.86$.
>
> Look again at the map. Measure the width of Greenland along the 75th parallel and the width of the Atlantic Ocean along the Equator (sixth horizontal, northern Brazil to Gabon at the western edge of Africa). On [my measurements of] the map, the ratio Greenland/Atlantic is about 0.61. On Earth, it is about 660 mi/4100 mi ≈ 0.16. By those numbers, the map overstates the ratio by 0.61/0.16 ≈ 3.8.

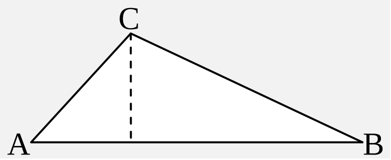
Mercator's fix was to exaggerate north-south distance to the same extent as the east-west. To do that, the band of the globe near latitude 1° had to stretch vertically about 1/cos 1°, the band near latitude 2° had to stretch about 1/cos 2°, …. (In the language of calculus: Mercator numerically integrated the function $1/\cos x = \sec x$.) Observe that the map's parallels of latitude (horizontal, also at multiples of 15°) get farther apart with distance from the Equator.

Therefore, there is still length magnification, increasing toward the top and bottom. Thus, northern Greenland is exaggerated (vertically) even more than southern Greenland. But at small scale, the horizontal and vertical magnifications are equal. Consequently, *proportion* ("aspect ratio") is preserved: An Earthly triangle of sides 2 mi, 3 mi, 4 mi is represented on the map by a similar triangle. That means *angles* are preserved. A map with that property is said to be **conformal**. The property is important to mariners. Indeed, "for the Use of Navigators" is part of the first Mercator map's title. Conformality implies that a straight line on the map corresponds to a path of constant heading (compass direction) on Earth. Thus, the map's line from Lisbon to where the Equator leaves Brazil, slanting down and left at around a 45° angle, depicts a path on Earth whose heading really is 45° south of west.

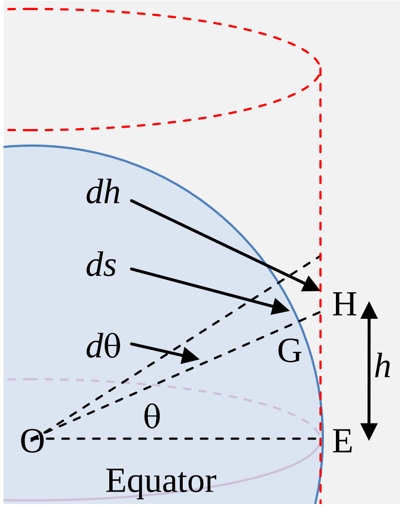
(Be careful to distinguish straight lines on the map from travel paths on the planet. Imagine on the map the line from Barrow to Nordkapp, from the top of Alaska to the top of Norway. It heads due east along the 71st parallel. Spot those places on a globe, though, and you see that the shortest flight path between them goes due north, over the North Pole, then due south.)

Exercises VI.A.2

1. (**Boyer**) a) Why did algebra and trigonometry develop more rapidly than geometry during the Renaissance?
   b) How do you account for the fact that many medieval and Renaissance mathematicians were clerics (like Oresme) and physicians (Rheticus)?
2. (**Boyer**) a) Find algebraically the other two sides of a triangle that has one side 5, the altitude to that side 3, and those remaining sides in the ratio √2:1.
   b) Given the solution, construct the triangle.
3. Prove the Law of Sines: In the figure at right, show that
   (sin A)/BC = (sin B)/AC.
   (Hint: Drop the altitude from C to side AB. The figure does not show the general case; it has angles A and B both acute. Draw the situation where, say, A is right or obtuse, and prove the same relation.)



4. (Calculus) Sometimes Mercator's method is incorrectly described as **cylindrical projection**. In the figure at left, an endless right circular cylinder (dashed red), whose axis and diameter match those of Earth (blue), hugs Earth. The point G at latitude θ on Earth's surface "projects" onto H, where the extension of OG meets the cylinder. After all the surface features are projected, a vertical cut along the cylinder opens it to a flat map.
   Show that this description does not match what Mercator did:
   a) Express the height *h* = HE in terms of θ and OE.
   b) Use (a) to find the derivative *dh*/*d*θ.
   c) Add the differential angle *d*θ to the latitude. This angle cuts off an arc *ds* long on the circle, and the arc projects to an added height *dh* on the cylinder. Show that the magnification *dh*/*ds* is $\sec^2\theta$. That is not the magnification sec θ required to fit the vertical stretching to the map's horizontal stretching.



Section VI.B. Algebra: The Cubic and Quartic Equations

The most dramatic algebraic development of the Renaissance was the solution of the cubic and quartic equations.

In 1545, Geronimo (Girolamo?) Cardano (1501-76) published *Ars Magna* ("The Great Art") with a partial solution of the cubic. The trouble was, it was not his solution. Tartaglia (an insult, meaning "stammerer," that the physician Niccolo Fontana chose to take as his name) had found it some years before—and Cardano said so—but had chosen to keep it secret. It took all Cardano could do to persuade Tartaglia (1500-1557) to confide it to him, under a vow of secrecy. But Tartaglia was not the discoverer, either. It had been the discovery of Scipione del Ferro (1465-1526), who had also kept it under wraps. Finding out about del Ferro, Cardano decided he was not bound by his promise. Read about the multiple imbroglios—and name-calling and challenges, plus much other important mathematics history—in Boyer, and most especially in Mario Livio's *The Equation That Couldn't Be Solved*.

1. Reducing the General Cubic

In the most general form

$$\alpha t^3 + \beta t^2 + \gamma t + \delta \ = \ 0$$

of third-degree equation, divide by the **leading** (highest degree) coefficient to make it look like

$$t^3 + Bt^2 + Ct + D \ = \ 0.$$

We then apply a transformation.

Theorem 1. The substitution $t = x - B/3$ always eliminates the quadratic (second-degree) term.

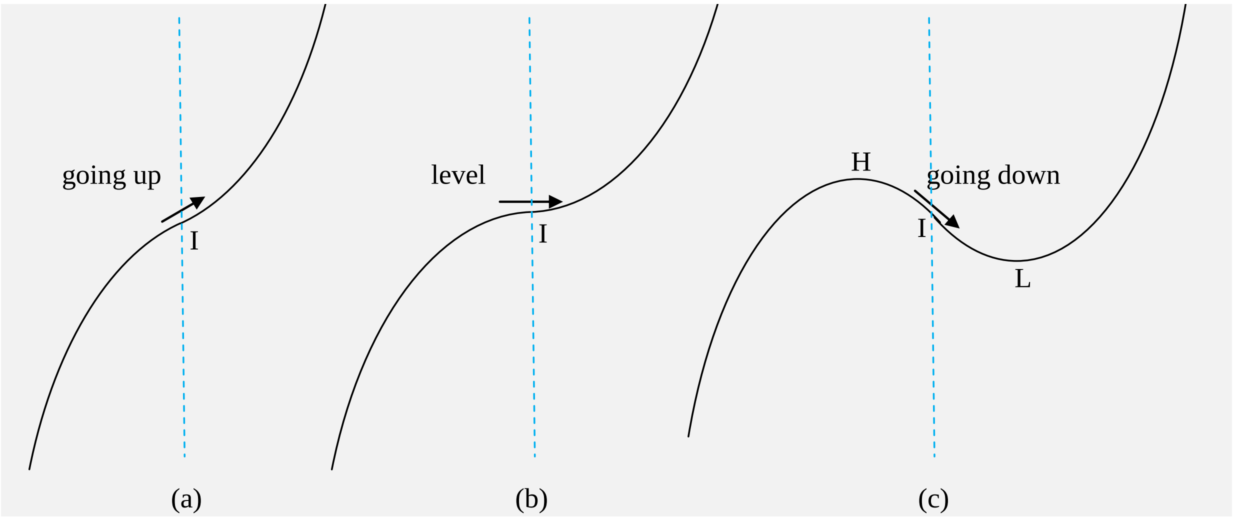
> Substituting $t = x - B/3$ turns the last form into
>
> $$(x - B/3)^3 + B(x - B/3)^2 + C(x - B/3) + D \ = \ 0,$$
>
> or
>
> $$x^3 - 3x^2B/3 + 3xB^2/9 - B^3/27 + Bx^2 - 2BxB/3 + BB^2/9 + Cx - BC/3 + D \ = \ 0.$$
>
> You can see that the only two quadratic terms (red) cancel. (Compare Exercise 1.) The substitution leaves us with an equation of the form
>
> $$x^3 + bx + c = 0.$$

Thereby the many possible cases of cubic equation (Exercise V.A.3:6) reduce to just three, just like the quadratic. (Remember that we need not bother if either $b = 0$ or $c = 0$.)

As modern types, we gain much by bringing coordinate geometry to this algebra. The graph of

$$y = t^3 + Bt^2 + Ct + D$$

is always an S-curve. There are three possible shapes of S, shown as solid curves in the figure below. In panel (a), the graph is going up to the right at the **inflection**, the point I where it changes from curving right to curving left. In panel (b), it momentarily levels off at I. In panel (c), it is going down. In that last case, there is necessarily a high point H to the left of I and a low point L to the right. For any of them,



the inflection is at the place where $t = -B/3$. For that reason, the substitution $x = t + B/3$ moves the inflection to the place where $x = 0$. In each of the panels, the dashed blue line is the y-axis for the graph of the transformed

$$y = x^3 + bx + c.$$

We do not show the x-axis, but it would be c below I. (See Exercise 2 for all of this.)

Exercises VI.B.1

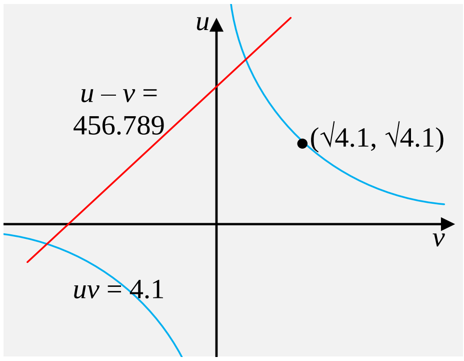
1. Use the substitution $t = x - 2/3$ to turn Fibonacci's cubic
$$t^3 + 2t^2 + 10t = 20$$
into the form
$$x^3 + bx + c = 0.$$

2. (Calculus) a) Show that the graph of
$$y = t^3 + Bt^2 + Ct + D$$
has an inflection point where $t = -B/3$.
b) Show that if $b > 0$, then the transformed graph
$$y = x^3 + bx + c$$
has the shape shown in panel (a) of the figure in this section. In particular, show that in this case, the graph crosses the x-axis exactly once (so that
$$x^3 + bx + c = 0$$
has exactly one solution).
c) Show that if $b = 0$, then the graph has the shape shown in panel (b).
d) Show that if $b < 0$, then the graph has the shape in panel (c), with a high point to the left and a low point to the right of the inflection.

2. Cardano's Solution in the Definite Case

We look first at the instances of
$$x^3 + bx + c = 0$$
in which the x-coefficient b is positive. Recall that in this case, the equation has exactly one real solution. (Exercise 2 above asks for calculus evidence, but observe that x^3 increases as x increases. That means $x^3 + bx + c$ also increases with x. Accordingly, if the polynomial is zero at some x, then it must be positive to the right and negative to the left. [Why *must* it be zero at some x?]) Cardano found the one solution by substituting $x = u - v$, subject to the condition that $uv = b/3$.

> You should immediately ask why such a complicated scheme should work. Just imagine that $b = 12.3$ and the solution is $x = 456.789$. How can you be sure that among the pairs that satisfy $uv = 12.3/3$, there is one that captures $u - v = 456.789$?
>
> Here we give a geometric answer; do Exercise 4 for algebraic evidence. Draw the u-v coordinate system as at right. The graph of $uv = 4.1$ is a hyperbola (blue) with branches in Quadrants I and III, asymptotic to the two axes. The graph of $u - v = 456.789$ is a line sloping up to the right, with u-intercept 456.789. It is clear that the line cannot miss the hyperbola. Indeed, it must cross both branches, one high to the right of the u-axis, the other way left below the v-axis. Those intersections give the required u and v.



Theorem 1. For the transformed cubic equation
$$x^3 + bx + c = 0$$
with $b > 0$, the substitution $x = u - b/3u$ always produces the solution.

First, just make $x = u - v$. That turns the equation into

$$(u - v)^3 + b(u - v) + c \ = \ 0.$$

Multiplying out yields

$$u^3 - 3u^2v + 3uv^2 - v^3 + bu - bv + c \ = \ u^3 - u(3uv - b) + v(3uv - b) - v^3 + c \ = \ 0.$$

We can reduce the clutter by making $3uv - b = 0$. Accordingly, we select $v = b/3u$.

We now have

$$u^3 - v^3 + c \ = \ u^3 - b^3/27u^3 + c \ = \ 0.$$

We do not want the unknown u in the denominator. Multiply by u^3 to get

$$u^6 + cu^3 - b^3/27 = 0.$$

If that does not seem like progress—trading a third-degree problem for one of degree six—observe that the last equation has degree six, but is actually quadratic. It has u^3 and $(u^3)^2$. The discriminant

$$c^2 - 4(-b^3/27) \ = \ c^2 + 4b^3/27$$

is clearly positive. Therefore we have the two solutions

$$u^3 \ = \ (-c \pm \sqrt{[c^2 + 4b^3/27]})/2 \ = \ -c/2 \pm \sqrt{[c^2/4 + b^3/27]}.$$

Cardano, like al-Khwarizmi, did not countenance negative solutions. He would have discarded the -√ choice, because it would have given a negative u^3 (consult Exercise 5). We share no similar inclination, but let us humor him for a minute. From u^3 with the +√, we proceed to

$$u = \sqrt[3]{-c/2 \ + \ \sqrt{c^2/4 + b^3/27}}.$$

That does not answer the question. We need to subtract $v = b/3u$. We have

$$b/3u = \frac{b}{3\sqrt[3]{-c/2 + \sqrt{c^2/4+b^3/27}}}\,.$$

Remarkably, we can move that root to the numerator by a sort of rationalization. Multiply numerator and denominator by $\sqrt[3]{-c/2 \ - \ \sqrt{c^2/4 + b^3/27}}$. The result is

$$b/3u = \frac{b\sqrt[3]{-c/2 - \sqrt{c^2/4+b^3/27}}}{3\sqrt[3]{(-c/2)^2 - [c^2/4 + b^3/27]}}\ .$$

In that denominator, everything cancels except $\sqrt[3]{-b^3} = -b$. (Verify the cancellation!) Hence

$$b/3u = -\sqrt[3]{-c/2 \ - \ \sqrt{c^2/4 + b^3/27}}\ .$$

The solution to the cubic is

$$x = \sqrt[3]{-c/2 \ + \ \sqrt{c^2/4 + b^3/27}} + \sqrt[3]{-c/2 \ - \ \sqrt{c^2/4 + b^3/27}}.$$

Let us agree to call that last the **cubic formula.** It is immediate from this form of it that either choice in the ±√ produces the same solution. (Compare Exercise 1.) It also follows that if c is positive, then the solution x is negative. In that case, the radicand

$$-c/2 \ + \ \sqrt{c^2/4 + b^3/27}$$

is a positive number, evidently with positive cube root. At the same time,

$$-c/2 \ - \ \sqrt{c^2/4 + b^3/27}$$

is a negative number with greater absolute value, has therefore a negative cube root of greater absolute value than the first one. Hence the two cube roots sum to a negative number. (Compare Exercise 2.)

Cardano and his algebraic forebears ignored the forms

$$x^3 + bx + c = 0 \qquad \text{and} \qquad x^2 + bx + c = 0$$

with positive coefficients, because clearly they have no positive solutions. The quadratic might not have negative solutions either, but not so the cubic. It necessarily has one—precisely one—negative solution. (For the calculus evidence, do Exercise 6. We will see algebraic evidence in about 90 years.)

If instead $c = -d$ is negative, then the equation is the case the algebraists would have written as

$$x^3 + bx = d.$$

The evidence from either algebra or calculus points to a positive solution, and the formula confirms that. Of the two radicands,

$$d/2 \ + \ \sqrt{d^2/4 + b^3/27}$$

is positive,

$$d/2 \ - \ \sqrt{d^2/4 + b^3/27}$$

is negative, and the former has greater absolute value. Therefore the sum of their cube roots is positive (Exercise 3a).

Exercises VI.B.2. In these, you need a scientific calculator to evaluate the cube roots.

1. a) Make Cardano's substitution $x = u – v$ with $uv = 6/3 = 2$ in the equation

   $$x^3 + 6x = 88$$

   and solve for u.

   b) Calculate both possibilities (±√) for u.

   c) Calculate the corresponding values $v = 2/u$, then the two values $u – v$.

   d) Check by substitution that (c) solves the equation.

2. For the equation

   $$x^3 + 3x + 14 = 0:$$

   a) Calculate $-7 + \sqrt{(-7)^2 + 3^3/27}$ and $-7 - \sqrt{(-7)^2 - 3^3/27}$.

   b) Calculate their cube roots, then add the roots.

   c) Check that (b) solves the equation.

3. In Exercise VI.B.1:1, we (you) transformed Fibonacci's cubic into

   $$y = x^3 + 26/3\, x – 704/27 = 0.$$

   a) Apply the cubic formula to express the (necessarily positive) solution.

   b) Calculate (a) and compare with the estimate in Exercise V.B.2:5d. (Remember: The solution in (a) is offset by 2/3 from the earlier one.)

4. Show algebraically that there exist solutions to the simultaneous system of equations

   $$u – v = 456.789, \qquad uv = 4.1.$$

5. If $b > 0$, why is $-c/2 – \sqrt{[c^2/4 + b^3/27]}$ necessarily negative, irrespective of the sign of c?

6. (Calculus) a) Use the intermediate-value theorem to show that

   $$123x^3 + 456x + 789 = 0$$

   has a negative solution.

   b) Use the derivative to show that it can have *only one* solution.

---

3. Cardano’s Limitation in the Indefinite Cases

The cases of

$$x^3 + bx + c = 0$$

in which b is negative need separate treatment. There is no better substitution than Cardano’s, which may or may not work. It might work to find the lone positive solution; it might work, when there is no positive solution, to find the lone negative solution; and it might find the negative one, even though there is a positive one. Instead, it might fail, unable to find any solution even when *we can tell* that there must be a positive one and two negatives, or vice-versa. (See Exercise 6.)

A crucial step in the algebra of the previous section was the solution of a sixth-degree equation by means of the quadratic formula. The discriminant of the quadratic was

$$\Delta = c^2 + 4b^3/27.$$

(It is typical to denote a discriminant by Δ, the Greek capital *delta*.) If Δ is positive—as was guaranteed before, because b was positive—then we continue to the solution of the cubic. If Δ, which we will call **the discriminant of the cubic**, is negative, then the solution process stops. Notice the important nature of that inference. We conclude that *the solution process fails*, not that the cubic has no solution.

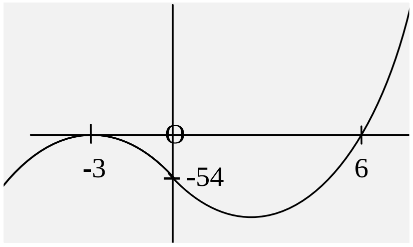
> In developing the formula to solve the quadratic
>
> $$ax^2 + bx + c = 0,$$
>
> we divide by a and complete the square to write
>
> $$(x + b/2a)^2 = -c/a + b^2/4a^2 = (b^2 - 4ac)/4a^2.$$
>
> This equation is **equivalent** to the original: It has the same solutions. If $b^2 - 4ac < 0$, then the last equation has no real solution. We justifiably conclude that the original has no real solution.
>
> When we apply the quadratic formula to the Cardano substitution, a negative Δ does not imply that the cubic has no solution. A cubic equation *must* have a real solution. Negative Δ simply blocks the Cardano approach to solving the cubic.

The discriminant being so important, we organize our survey—which we undertake with examples instead of a bunch of symbols—by the sign of Δ.

a) zero discriminant

> A cheap way to make $\Delta = 0$ is to take $b = -27$. (Taking $b = -3$ is a little too economical.) Then c has to be ±54. (Verify.)
>
> For $c = -54$, our equation is
>
> $$x^3 - 27x - 54 = 0.$$
>
> The cubic formula gives
>
> $$x = \sqrt[3]{-c/2} + \sqrt[3]{-c/2} = 6.$$
>
> That means $x - 6$ is a factor of the cubic, for a reason that we will see in the future. From there, the factoring is easy. The other factor has to include x^2 and 9. Thus,
>
> $$x^3 - 27x - 54 = (x - 6)(x^2 + ? + 9).$$
>
> The middle term in that second factor has to multiply by the first x to eliminate the $-6x^2$ term that the product already displays. We deduce
>
> $$x^3 - 27x - 54 = (x - 6)(x^2 + 6x + 9) = (x - 6)(x + 3)^2.$$
>
> That factorization is not an accident. If $\Delta = 0$, then $x^3 + bx + c$ has to be the product of the first-degree factor $\left(x - 2\sqrt[3]{-c/2}\right)$ and the perfect square $\left(x + \sqrt[3]{-c/2}\right)^2$ (Exercise 4).

From the factorization, we see that $x^3 - 27x - 54$ has a unique positive root at $x = 6$ and a unique negative one at $x = -3$. (We will mostly stick to the language “the equation *has a solution*” and “the polynomial *has a root*.”) However, because the factor $(x + 3)$ appears to the power 2, we say that $x = -3$ is a **double root** or a **root of multiplicity 2**. Under this usage, we say that the cubic has one positive root and “two negative roots.”



Now view the graph, in the figure at right. From the factoring, we can see that $x^3 - 27x - 54$ is positive rightward from $x = 6$ and negative leftward, except zero at $x = -3$. Knowing the general shape of such a cubic, we sketch at right the graph of

$$y = x^3 - 27x - 54.$$

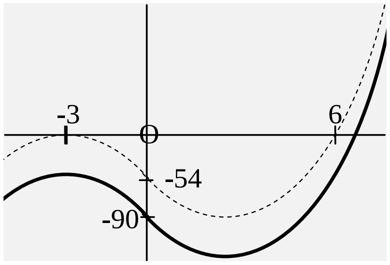
Notice that the graph is tangent to the x-axis at the double root.

In this situation, we have seen, Cardano’s method captures the **simple** root and not the double. It does the same if $c = +54$, even though then the captured root is negative (Exercise 1).

b) positive discriminant

To make $\Delta = c^2 + 4b^3/27$ positive, we can simply take the previous example and boost the absolute value of c. Accordingly, we examine

$$x^3 - 27x - 90 = 0.$$



The cubic formula gives

$$x = \sqrt[3]{45 + \sqrt{45^2 - 27^2}} + \sqrt[3]{45 - \sqrt{45^2 - 27^2}}$$
$$= \sqrt[3]{81} + \sqrt[3]{9} \approx 6.4.$$

Factoring is out of the question, but graphing is easy. The graph of

$y = x^3 - 27x - 54$ (previous figure)

looks like the dashed curve at right. Hence the current cubic

$$y = x^3 - 27x - 90 = (x^3 - 27x - 54) - 36$$

graphs as the solid curve. The negative root has disappeared, and the lone positive root is further to the right. (Compare Exercise 2, where the positive root disappears and there is a lone negative root.)

Thus, when $\Delta > 0$ with $b < 0$, Cardano’s method finds the one real root, on either side of zero.

c) negative discriminant

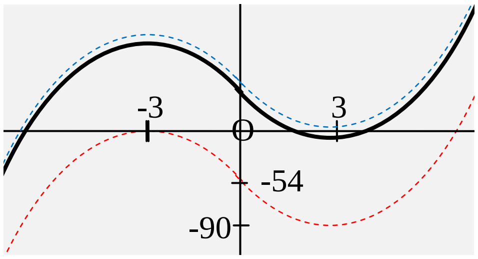
Stick with c as our guide. We can make Δ negative in our examples by reducing $|c|$. Examine

$$x^3 - 27x + 46 = 0.$$

For this one,

$$\Delta = 46^2 + 4(-27)^3/27 = -800.$$

We cannot complete Cardano’s process.



Consider the graph of

$$y = x^3 - 27x + 46.$$

Calculation gives $y = 46, 20, 0, -8$ when $x = 0, 1, 2, 3$, respectively. Without calculating, we see that y is high and positive when $x = 10$, low and negative when $x = -10$. The graph (heavy black in the figure at right) must cross the x-axis once between $x = -10$ and $x = 0$, again between $x = 0$ and $x = 3$ (we know where), and a third time between $x = 3$ and $x = 10$. It lies above the graph (dashed red) whose upper turning point is at the x-axis, and below the one (dashed blue) with lower turn at the x-axis.

In this case, the original equation has *three distinct roots*, yet Cardano’s method is blind to them all.

Revisit our question why the substitution $x = u - v$ should capture a solution if uv has to be $b/3$.

In VI.B.2, where b was positive, we saw that the uv-graph given by

$$u - v = (\text{the solution})$$

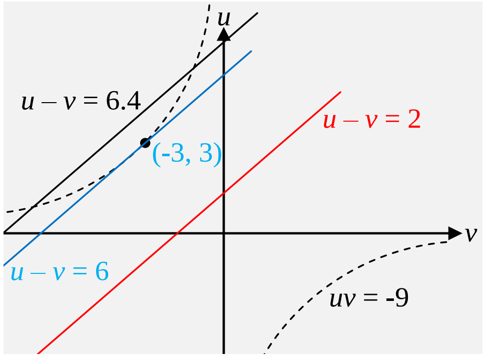
had to intersect the graph with $uv = b/3$. Now we have $b = -27$. The hyperbola

$$uv = b/3 = -9$$

(dashed curve at right) has its branches in Quadrants II and IV. You can see in the figure *three* possibilities: The line

$$u - v = \text{constant}$$

might cross the hyperbola (black line), just touch it (blue), or miss completely (red).



The borderline ("touch") case is precisely the boundary between the substitution's success and failure: Δ = 0. When we made

$$c^2 + 4b^3/27 = 0$$

by taking $c = -54$ (section (a)), we found the solution to be

$$x = u - v = 6.$$

The graphs of

$$u - v = 6 \qquad \text{and} \qquad uv = -9$$

meet only at $(v, u) = (-3, 3)$ (the figure and Exercise 3a). At that point, the hyperbola has a 45° inclination, same as the line; the two graphs are tangent. The substitution produces unique u and v.

When we took $c = -90$ (section (b)) to make Δ positive, we found the solution

$$x = u - v \approx 6.4.$$

The line $u - v = 6.4$ meets the hyperbola $uv = -9$ at two places (figure and Exercise 3b). Just as in VI.B.2, when the substitution produces two values of u, the resulting two values of $u - v$ match.

Finally, in our current setting of $\Delta < 0$ with $c = 46$, we see that $x = u - v = 2$ is a solution of the cubic equation. There are plenty of places where $u - v = 2$, but none simultaneously satisfies $uv = -9$ (figure and Exercise 3c). Therefore Cardano's substitution cannot find solutions to the cubic.

Exercises VI.B.3

1. For the equation

$$x^3 - 27x + 54 = 0:$$

a) What solution does the cubic formula give?
b) In view of (a), how does the polynomial factor?
c) What is the uncaptured, positive double root?
d) Sketch the graph of

$$y = x^3 - 27x + 54.$$

Check that this one is the graph in the text, lifted by 108; that is just enough to move the text's low point, from $y = -108$ when $x = 3$, up to the x-axis.

2. a) Apply the cubic formula to

$$x^3 - 27x + 90 = 0$$

to find a solution.
b) How does the graph of

$$y = x^3 - 27x + 90$$

compare with the one in Exercise 1? Use the graph to explain why the solution (a) is negative and left of $x = -6$.

3. Solve each (simultaneous) system:
   a) $u - v = 6$ and $uv = -9$
   b) $u - v = 6.4$ and $uv = -9$
   c) $u - v = 2$ and $uv = -9$.
4. a) Show that if $c^2/4 + b^3/27 = 0$, then
   $$x^3 + bx + c \quad = \quad \left(x - 2\sqrt[3]{-c/2}\right)\left(x^2 + 2\sqrt[3]{-c/2}\,x + \sqrt[3]{c^2/4}\right),$$
   and the last factor is a perfect square.
   b) Show that if $b^2 - 4ac = 0$, then
   $$ax^2 + bx + c \quad = \quad a(x + b/2a)^2.$$
   That makes $x = -b/2a$ a double root of the quadratic polynomial.
5. We know that the line
   $$u - v = \text{constant}$$
   might miss the hyperbola $uv = -9$. But
   $$u + v = \text{constant}$$
   slopes down to the right; it cannot possibly miss. Why do we not use the substitution
   $$x = u + v, \qquad \text{subject to the requirement } uv = -9?$$
6. (Calculus, although algebra will also serve) When b is negative, why is it impossible for
   $$x^3 + bx + c = 0,$$
   to have three positive roots, or three negative roots?

---

4. Bombelli's Solution

It took an engineer to show that the seeming failure of Cardano's method was, let us say, imaginary. In doing so, Rafael Bombelli (1526-1572) introduced a whole new world of numbers.

a) the quadratic puzzle

Cardano had considered this question: Find two numbers whose sum is 10 and whose product is 40.

To satisfy the first condition, represent the numbers by x and $10 - x$. To meet the second, write

$$x(10 - x) = 40, \qquad \text{and rewrite}$$
$$x^2 - 10x + 40 = 0.$$

The discriminant is $10^2 - 4(40) = -60$. We conclude that there is no solution. (See the graphical evidence in Exercise 1.)

Ignore the negative discriminant, and apply the quadratic formula anyway. Write

$$x = (10 \pm \sqrt{-60})/2.$$

Cardano "toyed" with such expressions (**Boyer**'s word), but called them "sophistic". Recall that this was a man who was unhappy when forced to use *negative* numbers, which he called "fictitious." Bombelli chose instead to treat them like numbers.

Assume they do work like numbers. Then their sum is

$$(10 + \sqrt{-60})/2 + (10 - \sqrt{-60})/2 \;=\; 20/2 \;=\; 10,$$

because they have a common denominator, the like terms 10 and 10 add up, and the like terms $\sqrt{-60}$ and $-\sqrt{-60}$ cancel. Also their product is

$$[(10 + \sqrt{-60})/2]\,[(10 - \sqrt{-60})/2] \;=\; [10^2 - (\sqrt{-60})^2]/4 \;=\; [100 - -60]/4 \;=\; 40,$$

because sum times difference is always the difference of squares, and squaring the square root gives the radicand. Therefore the two "numbers" do answer the question.

Recall now our rule (section IV.A.3): Before you get to call things "numbers," you must specify how to do their arithmetic. There is no problem with, say,

$$(10 + \sqrt{-60})/2 - (10 - \sqrt{-60})/2 \;=\; \sqrt{-60},$$

because the radicals are like. What about $\sqrt{-25} + \sqrt{-36}$? The answer is to make them somewhat like.

> These things are supposed to work like real numbers. That would mean
>
> $$\sqrt{-25} = \sqrt{25}\,\sqrt{-1} \qquad \text{and} \qquad \sqrt{-36} = \sqrt{36}\,\sqrt{-1}.$$
>
> Therefore we can at least combine
>
> $$\sqrt{-25} + \sqrt{-36} \;=\; 5\sqrt{-1} + 6\sqrt{-1} \;=\; 11\sqrt{-1}.$$
>
> Similarly
>
> $$\sqrt{-25} + \sqrt{-60} \;=\; (5 + \sqrt{60})\sqrt{-1};$$
>
> $5 + \sqrt{60}$ is a perfectly good real number. Consequently, we do not need to work with square roots of all negatives, just with $\sqrt{-1}$.

b) the arithmetic of complex numbers

Let us switch now to modern idiom. We represent $\sqrt{-1}$ by i, a notation that came 200 years after Bombelli. Then the expression $a + bi$ represents a **complex number**, a name from even later. This i is a creature, an entity, or finally "a number" characterized by two properties:

1. It is not a real number. Therefore we may not combine it with real numbers. Thus, $3 + i$ cannot combine into a single term, and $3 \times i$ is simply $3i$; and
2. We may multiply it by itself, and the product is -1.

Notice that i and $-i$ are unequal. If $i = -i$ were true, then addition would give us $2i = 0$. Those two numbers cannot be equal, because their squares are -4 and 0.

Complex addition, subtraction, and multiplication are straightforward.

> We do addition and subtraction according to like terms:
>
> $$(a + bi) \pm (c + di) \;=\; (a \pm c) + (b \pm d)i.$$
>
> Multiplication follows the distributive law:
>
> $$(a + bi)(c + di) \;=\; ac + adi + bci + bdi^2.$$
>
> The last appears to introduce a new element, but of course $i^2 = -1$. The product takes the right form,
>
> $$(a + bi)(c + di) \;=\; (ac - bd) + (ad + bc)i.$$

We do division by a kind of rationalization.

> (We should call it "real-ization.") Facing
>
> $$(a + bi)/(c + di),$$
>
> multiply numerator and denominator by $c - di$, then simplify:
>
> $$\frac{a+bi}{c+di}\,\frac{c-di}{c-di} \;=\; \frac{(ac+bd)+(bc-ad)i}{c^2-d^2i^2} \;=\; \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}\,i.$$
>
> That produces the needed form. Notice that the final fractions are legal; $c^2 + d^2$ is positive, unless c and d are both zero, in which case the original division was illegal. (See Exercise 2.)

Given $c + di$, $c - di$ is called its (**complex**) **conjugate**. (What is the conjugate of $c - di$?) It is universal to symbolize a complex number's conjugate by use of an **overbar**. Thus, if $z = c + di$, then

$$\bar{z} \;=\; \overline{c + di} \;=\; c - di.$$

One important property of conjugation is that it is compatible with the arithmetic operations. That is,

$$\overline{z + w} = \bar{z} + \bar{w}, \qquad \overline{z - w} = \bar{z} - \bar{w}, \qquad \overline{zw} = \bar{z}\,\bar{w}, \qquad \overline{z/w} = \bar{z}/\bar{w}.$$

In words, the conjugate of a sum is the sum of the conjugates, and similarly with difference, product, and quotient (Exercise 3).

Finally, there is some associated vocabulary. Given $z = c + di$, we call c the **real part** of z, d the **imaginary part** of z. If $d = 0$, we say $z = c$ **is real**; if instead $c = 0$, we say $z = di$ **is** (**purely**) **imaginary**. Notice that the imaginary part of *z = c + di is not imaginary*; it is the real number d. More important, do not describe a real number as "not complex." Every real $r = r + 0i$ is a complex number with zero imaginary part, just as a square is a rectangle of a special type.

Exercises VI.B.4b

1. a) Sketch the graph of
   $y = x(10 - x)$.
   b) Find the coordinates of the highest point on the graph, to determine that $x(10 - x)$ cannot reach as high as 40. (Calculus is not needed; algebra can answer.)
2. a) Evaluate $(2 + 3i) + (4 - 5i)$, $(2 + 3i) - (4 - 5i)$, $(2 + 3i)(4 - 5i)$, $(2 + 3i)/(4 - 5i)$.
   b) Verify that your quotient times $(4 - 5i)$ gives $(2 + 3i)$.
3. a) For the sum, difference, and product in Exercise 3a, show that the conjugate of the result is the sum, difference, or product of the conjugates.
   b) Show that the conjugate of a cube is the cube of the conjugate.

c) extending Cardano

Bombelli's book *Algebra* presented the rules for operating with the new numbers. With the rules, he gave a way to interpret the cubic formula when it "failed" to work.

View our example cubic,

$x^3 - 27x + 46 = 0.$

Accepting complex answers, we take from the cubic formula

$$x = \sqrt[3]{-23 + \sqrt{-200}} + \sqrt[3]{-23 - \sqrt{-200}}.$$

We will apply Bombelli's reasoning to this example and leave to Exercise 2 the actual cubic he took up from Cardano.

Write $(-23 + \sqrt{-200})$ as $(-23 + 10i\sqrt{2})$. (Writing $10i\sqrt{2}$ is better than $10\sqrt{2}i$, because the latter looks too much like $10\sqrt{[2i]}$.) Its cube root would be a complex $(a + bi)$ with

$$\begin{aligned} -23 + 10i\sqrt{2} &= (a + bi)^3 \\ &= a^3 + 3a^2bi + 3a(bi)^2 + (bi)^3 \\ &= (a^3 - 3ab^2) + i(3a^2b - b^3). \qquad \text{(Verify!)} \end{aligned}$$

To find the cube root, we need only solve the simultaneous equations

$$\begin{aligned} -23 &= a^3 - 3ab^2 \\ 10\sqrt{2} &= 3a^2b - b^3. \end{aligned}$$

We have made the trouble considerably worse. We need a different way.

Bombelli made two observations. One was that

$-23 + \sqrt{-200} = -23 + 10i\sqrt{2}$ and $-23 - \sqrt{-200} = -23 - 10i\sqrt{2}$

are conjugates. It follows that their cube roots are conjugates, because by Exercise 3b above, if

$-23 + 10i\sqrt{2}$ is the cube $(a + bi)^3$,

then its conjugate

$-23 - 10i\sqrt{2}$ is the cube $(a - bi)^3$.

Second, and more important, he already knew that the cube roots add up to the solution $x = 2$. From

$(a + bi) + (a - bi) = 2,$

he deduced $a = 1$.

Now the hunt for the cube roots is half-done. We want

$$-23 + 10i\sqrt{2} \;=\; (1 + bi)^3 \;=\; (1 - 3b^2) + i(3b - b^3).$$

(Verify that this result comes from the line marked "(Verify!).") For that to happen, we need

$$-23 = 1 - 3b^2.$$

That relation allows

$$b \;=\; \pm\sqrt{8} \;=\; \pm 2\sqrt{2}.$$

Simultaneously, we need

$$10\sqrt{2} = 3b - b^3.$$

That one forces $b = -2\sqrt{2}$. (Use Exercise 3a as a check.) We therefore have

$$x \;=\; \sqrt[3]{-23+\sqrt{-200}} + \sqrt[3]{-23-\sqrt{-200}} \;=\; (1 - 2i\sqrt{2}) + (1 + 2i\sqrt{2}) \;=\; 2.$$

Bombelli's method, operating on these new numbers, produces the old solution to the cubic.

d) extending Bombelli

Of the cube roots of $(-23 + 10i\sqrt{2})$—the complex numbers $(a + bi)$ satisfying the equation

$$-23 + 10i\sqrt{2} = (a + bi)^3\text{—}$$

Bombelli found the one that has real part $a = 1$. We saw in (c) that this one equation is equivalent to the real-number system

$$\begin{aligned} -23 &= a^3 - 3ab^2 \\ 10\sqrt{2} &= 3a^2b - b^3. \end{aligned}$$

Bombelli could not solve it, and we are not about to try. Nevertheless, the degree of the system suggests that there might be other answers. We will eventually see that every nonzero complex number has two distinct square roots, three cube roots, four fourth roots, …. In fact, we will learn to approximate them. Here we simply state the other cube roots that those future methods will estimate (or even evaluate):

$$z = (-1/2 + \sqrt{6}) + i(\sqrt{2} + \sqrt{3}/2) \qquad \text{and} \qquad w = (-1/2 - \sqrt{6}) + i(\sqrt{2} - \sqrt{3}/2).$$

With our current knowledge, we cannot determine those roots. Checking them, however, is just complicated arithmetic. The real part of z^3 is

$$\begin{aligned} &[-1/2 + \sqrt{6}]^3 - 3(-1/2 + \sqrt{6})\,(\sqrt{2} + \sqrt{3}/2)^2 \\ &\quad= [-1/8 + 3\sqrt{6}/4 - 18/2 + 6\sqrt{6}] + (3/2 - 3\sqrt{6})\,(11/4 + \sqrt{6}) \\ &\quad= (-1/8 - 9 + 33/8 - 18) + \sqrt{6}(3/4 + 6 + 3/2 - 33/4) \\ &\quad= -23 + 0\sqrt{6} \quad = \quad -23. \end{aligned}$$

The imaginary part is

$$\begin{aligned} &3\,(-1/2 + \sqrt{6})^2\,(\sqrt{2} + \sqrt{3}/2) - [\sqrt{2} + \sqrt{3}/2]^3 \\ &\quad= (25/4 - \sqrt{6})\,(3\sqrt{2} + 3\sqrt{3}/2) - [2\sqrt{2} + 3\sqrt{3} + 9\sqrt{2}/4 + 3\sqrt{3}/8] \\ &\quad= \sqrt{2}(75/4 - 17/4) \;+\; \sqrt{3}(75/8 - 27/8) \;+\; \sqrt{12}(-3) \;+\; \sqrt{18}(-3/2) \\ &\quad= 58\sqrt{2}/4 \;+\; 6\sqrt{3} \;+\; 2\sqrt{3}(-3) \;+\; 3\sqrt{2}(-3/2) \quad = \quad 10\sqrt{2}. \end{aligned}$$

Verifying w^3 is the same arithmetic and just as much fun; we skip it. But consider Exercise 2d.

Refocus on the example

$$x^3 - 27x + 46 = 0.$$

Substitute z and its conjugate for the cube roots in the cubic formula, to write

$$x \;=\; z + \bar{z} \;=\; 2(-1/2 + \sqrt{6}).$$

That gives us the solution $x = -1 + 2\sqrt{6}$, which you can check in Exercise 3b. Using w instead, we get

$$x \;=\; w + \bar{w} \;=\; 2(-1/2 - \sqrt{6}).$$

That spots the third solution, $x = -1 - 2\sqrt{6}$ (Exercise 3c). Cardano's method is complete (it always gives all the solutions) when you work with Bombelli's newfangled numbers.

[**Struik** (p. 92) makes an interesting point. In teaching algebra, we typically introduce complex numbers to treat quadratics with no real solutions. This is striking, given that the numbers arose in the treatment of cubics that were *known* to have real solutions.]

Exercises VI.B.4d

1. (after **Boyer**) In what specific way did the solution of the cubic lead to the development of complex numbers?
2. For the cubic equation
   $x^3 = 15x + 4$:
   a) Write the solution given by the cubic formula.
   b) Show that the two cube roots in (a) can be given by $2 \pm i$.
   c) What solution does (b) imply? Check that it solves the equation.
   d) Show that the two cube roots in (a) can also be given by $(-1 - [\sqrt{3}]/2) \pm (-1/2 + \sqrt{3})i$.
   [Hint: Skip this, unless you have excess spare time. But do (e).]
   e) What solution does (d) imply? Check that this number also solves the equation.
3. a) Verify that
   $(1 - 2i\sqrt{2})^3 = -23 + 10i\sqrt{2}$.
   b) Verify that $(-1 + 2\sqrt{6})$ solves $x^3 - 27x + 46 = 0$.
   c) Verify that $(-1 - 2\sqrt{6})$ also solves $x^3 - 27x + 46 = 0$. (Is there a shortcut?)

5. Ferrari and the Quartic Equation

Lodovico (or Luigi) Ferrari produced a method for solving equations of fourth degree. It had general elements in common with the method for the cubic. It began with a simple translation to eliminate one term. (For the cubic, it had been $x = t + b/3$ to eliminate the quadratic term; for the quartic, $x = t + b/4$ eliminates the cubic term.) It needed to consider separate cases. It used manipulations, introduction of a second variable, and further transformations to make the solution come down to solving a cubic equation. We will not describe his approach, even in words. A verbal description including some algebra is in **Boyer**, and a well-detailed algebraic treatment is on Wikipedia®.

It makes sense that Ferrari (1522-1565) reduced the quartic question down to cubic size, just as Cardano made solving the cubic come down to solving a quadratic. The odd thing is that the latter came later. Ferrari's solution, which Cardano identified as such, was ready before the *Ars Magna* came out. It had to wait for the solution of the cubic to be published.

Section VI.C. Viète and the Evolution of "Algebra"

Recall that the earliest mathematical thought we could call "algebra" was tied to practical problems about the distribution of foods, inheritances, land, and the like. Its methods were specific, sometimes geometric. In the hands of al-Khwarizmi, it retained its practical use, but became oriented toward general methods for solution of equations. With the progression to cubic and quartic equations, it became less attached to everyday problems; the solution methods were all that mattered.

You can see that the obvious next step would be solution of the quintic. With that, algebra would separate further from questions related to physical objects. However, we do not have to go that far. Already in Bombelli's willingness to deal with complex numbers—which you could not represent even by *directed* lengths, as you can do with negative numbers—we have a completely new outlook. It would be three centuries before a physical use, the analysis of electrical oscillations, would appear for complex numbers.

Emblematic of the emerging view of what constituted algebra was François Viète (1540-1603). Viète foreshadowed what was to come in a number of ways. For one thing, he was a Frenchman. For centuries, European mathematics had come largely from the Italian cities. (Remember that there was no “Italy” until 1871.) Venice was the port of entry for much of Arabic knowledge. Leonardo was from Pisa. Even more remarkable was the University of Bologna. Both del Ferro and Cardano taught there. So later would Cavalieri. Copernicus studied there. In the 1600’s, France would become preeminent in the development of mathematics. For another thing, Viète was a lawyer. Training in the law would begin to yield mathematicians, the way training for the church or medicine once did. (See Exercise 1.)

1. Parameters vs. Variables

Recall that Jordanus had used letters to represent quantities. Viète made extensive use of letters, but one thing he did was new and important. He used vowels for what we call “variables” and consonants for what we call “coefficients.” This achieved a separation of variables from “parameters.”

(In ordinary discourse, “parameter” often means “limit,” in the sense of boundary or restriction. The parameters of a contract, for example, limit what each side is allowed to do. In math, **parameters** are unspecified numbers whose assignment decides which specific example of something is at hand.)

The distinction made it possible to talk about or work with general categories of algebraic objects. We had such discussion, albeit with different letters. For example, we looked at “the general quadratic equation” or “general form of the quadratic equation”

$$ax^2 + bx + c = 0.$$

This (“three-parameter” form) is a simple, compact way to symbolize the whole class of quadratic equations. In our use, we followed the custom to “let *x* be the unknown.” The parameters *a*, *b*, and *c* were understood to be unspecified fixed numbers, whose specification would yield a particular equation.

Since the parameters of a general equation fix what specific equation is under study, they decide the solutions. Evidently, then, they must carry all the information about the equation. Thus, in the general quadratic form, the specification $a = 2$, $b = 3$, $c = 4$ gives

$$b^2 - 4ac \ = \ 9 - 4(2)(4) \ < \ 0.$$

That tells us the corresponding quadratic equation

$$2x^2 + 3x + 4 = 0$$

has no real solution. Similarly, for our (“two-parameter”) transformed cubic

$$x^3 + bx + c = 0,$$

we can tell from the sign of $c^2 + 4b^3/27$ how many real solutions there are, even if we cannot find them.

2. Relationships Between Solutions and Coefficients

The coefficients carry all the information about the equation. Hence they must somehow be related to the solutions. Viète was first to write about those relationships.

> Look at the quadratic case first, because then we can explicitly write the solutions. If
>
> $$r \ = \ (-b + \sqrt{[b^2 - 4ac]})/2a \qquad \text{and} \qquad s \ = \ (-b - \sqrt{[b^2 - 4ac]})/2a,$$
>
> then we can see that the sum and product of those solutions are
>
> $$r + s \ = \ -2b/2a \ = \ -b/a \qquad \text{and} \qquad rs \ = \ (b^2 - [b^2 - 4ac])/4a^2 \ = \ c/a.$$
>
> Notice that those relations hold even if the two solutions are complex, which the coefficients are not. They even hold if the “two” solutions are one (double root of the polynomial; see Exercise 2b).

The situation is more interesting for our cubic.

Theorem 1. If *u* and *v* are two distinct solutions of

$$x^3 + bx + c = 0,$$

then

$$u^2 + uv + v^2 = -b \qquad \text{and} \qquad uv^2 + u^2v = c.$$

Viète was old-fashioned enough to confine his attention to positive coefficients. His form would have been $x^3 + c = dx$, with positive *c* and *d*. He also eschewed negative solutions, never mind complex ones. But we, as usual, take on all comers as long as *b* and *c* are nonzero.

> To say that *u* and *v* are solutions is to say that
>
> $$u^3 + bu + c = 0 \qquad \text{and} \qquad v^3 + bv + c = 0.$$
>
> Subtract and factor to get
>
> $$u^3 - v^3 + bu - bv \;=\; (u - v)(u^2 + uv + v^2) + b(u - v) \;=\; 0.$$
>
> Because $(u - v)$ is not zero, we may legally divide by it to get the first relation. (See also Exercises 3 and 2b.)
>
> The subtraction eliminated *c*. Eliminate *b* by writing
>
> $$u^3 + c = -bu \qquad \text{and} \qquad v^3 + c = -bv,$$
>
> then dividing. From
>
> $$(u^3 + c)/(v^3 + c) \;=\; -bu/-bv \;=\; u/v \qquad \text{(Why are those divisions legal?),}$$
>
> we cross-multiply to get
>
> $$u^3v + cv \;=\; uv^3 + cu.$$
>
> Rearrange and factor to find
>
> $$\begin{aligned} u^3v - uv^3 \;&=\; uv(u^2 - v^2) \;=\; uv(u + v)(u - v) \\ &=\; cu - cv \;=\; c(u - v). \end{aligned}$$
>
> The second relation follows.

Notice that in this argument, we did not employ any knowledge of what the solutions are or how to find them. For that matter, we did not so much as evince *interest* in finding them. Viète was first to separate algebraic thought into what we might call three phases. The first is the part in which a problem or verbal description is turned into one or more equations. Thus, we might turn al-Khwarizmi’s “a square and four times its side sum to 60” to

$$x^2 + 4x = 60.$$

Second is reasoning about how algebraic quantities are related. We did that just above to relate the (unstated) roots. We had done it before with, for example, arguing why you can eliminate the square term in a general cubic equation. This phase is important for more than its equation-solving value. It adds to algebra a deductive component that had not been there in, say, al-Khwarizmi, and that moved it closer to the methods of geometry. Third phase is the algorithmic part, in which we do whatever it takes to produce those values of the unknown that make the equation true.

Still, we could use the relations—let us call them **Viète‘s equations**—to find the remaining solutions, provided we know one of them. (Check out Exercise 4.)

> Go back to our
>
> $$x^3 - 27x + 46 = 0 \qquad \text{(sections VI.B.4c and d).}$$
>
> We saw immediately that $u = 2$ is one solution. Any other solution *v* has to satisfy
>
> $$2^2 + 2v + v^2 = 27 \qquad \text{and} \qquad 2v^2 + 2^2v = 46.$$
>
> The two are equivalent equations. By the quadratic formula, they yield
>
> $$v = (-2 \pm \surd[4 + 92])/2 = (-2 \pm 4\surd 6)/2 = -1 \pm 2\surd 6.$$
>
> Those are the solutions we produced from thin air in VI.B.4d, then verified in Exercise 3 there.

Exercises VI.C.2

1. Why would the law become a source of mathematicians in the seventeenth century, but not before?
2. Assume that r and s are two distinct solutions of

   $ax^2 + bx + c = 0.$

   a) Mimic the argument from Theorem 1 to show, without explicitly writing the values of r and s, that they satisfy

   $r + s = -b/a$ and $rs = c/a.$

   b) Suppose $r = s$ is a double root of the quadratic. Do

   $r + s = -b/a$ and $rs = c/a$

   still hold? (Hint: Exercise VI.B.3:4b.)
3. Do Viète's equations remain valid if $u = v$ is a double root of

   $x^3 + bx + c = 0$?

   (Hint: Exercise VI.B.3:4a.)
4. In Exercise VI.B.4d:2, we met Cardano's cubic

   $x^3 = 15x + 4,$

   for which $x = 4$ is evidently a solution. Use Viète's equations to find the other two solutions. Match them against the one solution given back there.

3. Symmetry of the Relations

Of the relations between coefficients and solutions, one feature turned out two centuries later to be enormously important. It is that the relations are *symmetric* with respect to the solutions.

> For the two solutions r and s of the quadratic equation
>
> $ax^2 + bx + c = 0,$
>
> we wrote
>
> $r + s = -b/a$ and $rs = c/a.$
>
> Observe that if we switch the order of r and s, then we write the equivalent expressions
>
> $s + r = -b/a$ and $sr = c/a.$
>
> The relations are **symmetric in *r* and *s*.**

Sometimes the sum $r + s$ is called the sum of **the products of the solutions taken one at a time**. In strictly algebraic language, it would be the sum of **the first-degree terms** in the solutions. The names extend to the sum of the products of the solutions taken two at a time, or **second-degree terms**.

> The second-degree term rs is simply the one that uses both solutions, without repetition. If we allow repetition, then r^2 and s^2 are the other two-at-a-time products. The sum of all of them is
>
> $$\underline{\underline{r^2 + rs + s^2}} = [r + s]^2 - [rs].$$
>
> The double-underlined expression is symmetric, and both quantities in brackets are expressible in terms of a, b, and c (Exercise 1a).
>
> The sum of the third-degree products (where repetition is inevitable) is
>
> $$\begin{aligned}\underline{\underline{r^3 + r^2s + rs^2 + s^3}} &= r^2[r + s] + s^2[r + s] \\ &= (r^2 + s^2)[r + s] = ([r + s]^2 - 2[rs])[r + s],\end{aligned}$$
>
> again symmetric and expressible in terms of the coefficients (Exercise 1c).

For roots of our cubic polynomials, we already have Viète’s equations. Those, however, relate the roots two at a time, symmetrically. Our type of cubic will have three roots: three simple real solutions, or one simple real and one double real, or a simple real and two non-real complex solutions. Let us therefore look at the solutions u, v, w of

$$x^3 + bx + c = 0.$$

> We may assume $v \neq w$. Viète’s first equation gives
>
> $$-b = u^2 + uv + v^2 \qquad \text{and} \qquad -b = u^2 + uw + w^2,$$
>
> even if u matches either of v or w. (Consult Exercise VI.C.2:3.) Subtracting, we have
>
> $$0 \;=\; uv - uw + v^2 - w^2 \;=\; u(v - w) + (v + w)(v - w).$$
>
> Legal division by $(v - w)$ gives us
>
> $$\underline{\underline{u + v + w}} = 0,$$
>
> symmetric as always. The sum of the one-at-a-time products is zero.
>
> From there, we proceed to
>
> $$0 \;=\; 2(u + v + w)^2 \;=\; 2u^2 + 2v^2 + 2w^2 + 4uv + 4uw + 4vw$$
> $$= \;(u^2 + uv + v^2) + (u^2 + uw + w^2) + (v^2 + vw + w^2) + 3uv + 3uw + 3vw.$$
>
> By the first Viète equation, all three expressions in parentheses equal $-b$. We conclude
>
> $$0 \;=\; -3b + 3(uv + uw + vw).$$
>
> The sum of the products of the solutions, two at a time without repeats, is
>
> $$\underline{\underline{uv + uw + vw}} = b.$$
>
> That relation, together with Exercise 2a, gives us the sum of the two-at-a-time products.
>
> From the equation in red, we see that
>
> $$\underline{\underline{u^2 + v^2 + w^2}} = -2(uv + uw + vw) = -2b.$$
>
> (Do Exercise 2a.)
>
> For the products of degree 3, write
>
> $$0 \;=\; (u + v + w)^3 \;=\; u^3 + v^3 + w^3 + 3u^2v + 3uv^2 + 3u^2w + 3uw^2 + 3v^2w + 3vw^2 + 6uvw.$$
>
> (Try to see why that is true *without* actually multiplying out.) We get the single-variable products (the cubes) from the original equation (the cubic) as follows: Because u, v, w are solutions, we have
>
> $$u^3 + bu + c = 0,$$
> $$v^3 + bv + c = 0,$$
> $$w^3 + bw + c = 0;$$
>
> add them to write
>
> $$u^3 + v^3 + w^3 + b(u + v + w) + 3c \;=\; u^3 + v^3 + w^3 + 3c \;=\; 0,$$
>
> and we find
>
> $$\underline{\underline{u^3 + v^3 + w^3}} \;=\; -3c.$$
>
> We get the two-solution products by the second Viète equation:
>
> $$(3u^2v + 3uv^2) + (3u^2w + 3uw^2) + (3v^2w + 3vw^2) \;=\; 3c + 3c + 3c.$$
>
> Finally, for the three-solution product uvw, the equation in blue gives
>
> $$0 = -3c + 9c + 6uvw, \qquad \text{or}$$
> $$\underline{\underline{uvw}} = -c.$$
>
> Put it all together in Exercise 2b.

That is a long sequence from the deductive, theorem-proving phase. Notice that all the double-underlined expressions are symmetric in the roots. They are also **homogeneous**: All the terms in each expression are of the same degree. Homogeneity was important to Viète. [I must confess that I do not understand why. His predecessors had no objection to adding “a square and four times its side”—an area and a length—and Descartes would soon drop forever any requirement for homogeneity.]

Exercises VI.C.3

1. Let r and s be the solutions of $ax^2 + bx + c = 0$. Express in terms of a, b, and c:
   a) $r^2 + rs + s^2$
   b) $r^2 + s^2$
   c) $r^3 + r^2s + rs^2 + s^3$
   d) $r^3 + s^3$.
2. Let u, v, and w be solutions of $x^3 + bx + c = 0$, with $v \neq w$. Express in terms of b and c:
   a) the sum
   $u^2 + v^2 + w^2 + uv + uw + vw$
   of the second-degree products of the solutions.
   b) the sum
   $u^3 + v^3 + w^3 + u^2v + uv^2 + u^2w + uw^2 + v^2w + vw^2 + uvw$
   of the third-degree products.

4. Solution Via Trigonometry

Viète made important contributions to trigonometry in general. His work with the trigonometric functions gave trigonometry a deductive flavor, leading to relations divorced from the solution of triangles. The relations proved to be surprisingly applicable to the solution of equations.

a) one example

Recall the angle-sum formula

$$\cos (r + s) = \cos r \cos s - \sin r \sin s.$$

It yields the double-angle formula

$$\cos 2r = \cos^2 r - \sin^2 r = \cos^2 r - (1 - \cos^2 r) = 2 \cos^2 r - 1.$$

Apply it one more time, to get a **triple-angle formula**:

$$\begin{aligned} \cos 3r &= \cos 2r \cos r - \sin 2r \sin r \\ &= (2 \cos^2 r - 1) \cos r - 2 \sin r \cos r \sin r \\ &= 2 \cos^3 r - \cos r - 2 \cos r (1 - \cos^2 r) \\ &= 4 \cos^3 r - 3 \cos r. \end{aligned}$$

In this formula, set $r = 20°$. The substitution yields

$$1/2 = \cos 60° = 4 \cos^3 20° - 3 \cos 20°.$$

Multiply by 2 and move everything to one side, and you see that $x = \cos 20° \approx 0.94$ is a solution of

$$8x^3 - 6x - 1 = 0.$$

(See Exercise 2b.)

b) many examples

This use of the triple-angle formula may seem like a lot of work just to solve one cubic. However, Viète showed that *every cubic equation of the bad type* can be molded into the formula and solved by means of trigonometry.

Recall that the "bad type" is the equation

$$x^3 + bx + c = 0$$

in which the discriminant $c^2 + 4b^3/27$ is negative. Cardano's method hit a roadblock, and Bombelli wrote the solutions in terms of complex expressions he could not evaluate. Viète made the substitution

$$x = u \cos v.$$

(Necessarily, an infinity of combinations of u and v will give whatever value x has.)

The substitution turns the equation into

$$u^3 \cos^3 v + bu \cos v \;=\; -c.$$

To start making it resemble the formula, multiply by $4/u^3$:

$$4 \cos^3 v + 4b/u^2 \cos v \;=\; -4c/u^3.$$

Now we need the coefficient $4b/u^2$ to be -3. That forces

$$u \;=\; \pm\sqrt{}(4[-b]/3) \;=\; \pm 2\sqrt{}([-b]/3).$$

(The minus sign within the radical is best attached to the b; for the discriminant to be negative, b has to be negative.) Take the positive choice for now. With that u, our equation becomes

$$4 \cos^3 v - 3 \cos v \;=\; -4c/(2\sqrt{}[-b/3])^3 \;=\; -c\sqrt{}27)/(2[-b]^{3/2}]. \qquad \text{(Verify the last equality.)}$$

The expression on the left is undeniably cos $3v$. For the method to succeed, the fraction on the right has to be the cosine of *something*. For the fraction to be a cosine, it has to be between -1 and 1, inclusive. That means its square has to be at most 1. The square is $27c^2/(4[-b]^3)$. That fraction is (strictly) less than 1, because the numerator is smaller than the denominator:

$$0 < 27c^2 < -4b^3 \qquad \text{follows from} \qquad c^2 + 4b^3/27 < 0.$$

Therefore $x = u \cos v$ solves the equation as long as

$$u = \sqrt{}(-4b/3) \qquad \text{and} \qquad v \text{ is any angle with } \cos 3v = -4c/u^3.$$

Reconsider our example

$$x^3 - 27x + 46 = 0.$$

The method calls for

$$u \;=\; \sqrt{}(108/3) \;=\; 6 \qquad \text{and} \qquad \cos 3v \;=\; -184/6^3 \;\approx\; -.8519.$$

One angle answering to that cosine is $3v \approx 148.42°$, from which $v \approx 49.47°$. One solution is then

$$x \;=\; u \cos v \;=\; 6 \cos 49.47° \;\approx\; 3.90.$$

That matches the solution $x = -1 + 2\sqrt{}6$ from VI.B.4d.

(These numbers are from a scientific calculator. Viète, lacking such a tool, could nevertheless have worked at 0.01° precision. He published tables of all the trigonometric functions with 1-minute precision. A minute is 0.01666…°. Interpolating to 3/5 of that would have been easy. [Using minutes was a concession to tradition, plus astronomy; Viète himself always espoused calculation in *decimal* fractions.] Separately, if you think of that earlier symbolic solution as “exact” and 3.90 as only approximate, remember that our current solution also has exact, symbolic form:

$$x \;=\; \sqrt{}(-4\,[-27]/3) \cos (1/3 \cos^{-1} [-4\times 46/6^3]).)$$

If we had chosen the negative u, $u = -\sqrt{}(-4b/3) = -6$, then $\cos 3v = -4c/u^3$ would have taken the opposite sign: $\cos 3v \approx +.8519$. That would give $3v \approx 31.58°$, then $v \approx 10.53°$. Notice that cos v *does not change sign*; you cannot predict the sign of cos v from that of cos $3v$. Now we have the solution

$$x \;=\; u \cos v \;=\; -6 \cos 10.53° \;\approx\; -5.90.$$

That one matches the earlier solution $-1 - 2\sqrt{}6$.

What about the solution we saw by inspection, $x = 2$? Go back to

$$\cos 3v \approx -.8519.$$

The angle 148.42° = 180° – 31.58° is not the only one with that cosine. There is also

$$180° + 31.58° \;=\; 211.58°.$$

(Treat others in Exercise 1.) With that value of $3v$, we have $v = 70.53°$ and

$$x \;=\; u \cos v \;=\; 6 \cos 70.53° \;\approx\; 2.00.$$

Viète’s substitution finds all three solutions trigonometrically, *without venturing into complex numbers*. (See Exercise 2c.)

Exercises VI.C.4

1. We found cos $3v \approx -.8519$ for two angles $3v$, namely 148.42° and 211.58°.
   a) Find four other angles with the same cosine.
   b) Show that for each of them, cos v is one of the three values we already found:
   cos 10.53° or cos 49.47° or cos 70.53°. (You can avoid calculating the cosines by matching up the reference angles.)
2. a) Justify this statement:
   $8x^3 - 6x - 1 = 0$
   has three distinct real solutions.
   b) The polynomial has value -1 when $x = 0$ and +1 when $x = 1$. Therefore it must have a root between $x = 0$ and $x = 1$. Use calculation to approximate it to within 0.01. How does your approximation compare with cos 20°, which we know to be the one positive root?
   c) Find trigonometrically the other two real solutions.
3. Solve trigonometrically Cardano’s
   $x^3 = 15x + 4$.
   Match the solutions against Exercise VI.C.2:4.
4. a) Use cos $4r$ = cos $(2\times 2r)$ to show that
   $\cos 4r = 8cos^4 r - 8cos^2 r + 1$.
   b) Use trigonometry to find all solutions of
   $-1/2 = 8x^4 - 8x^2 + 1$.
   c) Solve the same equation using the quadratic formula. Do the answers agree with (b)?
5. We assumed in this subsection that the discriminant is negative. Does Viète’s substitution still give the solutions in the case where the discriminant is zero?
   a) Try it with the example
   $x^3 - 27x - 54 = 0$,
   which we solved in VI.B.3a.
   b) Show that it works in general. (Refer to Exercise VI.B.3:4a.)

5. Trigonometric Formulas

Viète’s development of trigonometry outside of triangles took two other surprising turns. One was in calculation, the other in multiple-angle formulas.

a) calculation

Viète showed how to use precise tables, like his, of trigonometric functions to turn multiplications and divisions into additions and subtractions. His discovery came just ahead of the discovery of logarithms, and both had the same purpose: to facilitate calculations needed in *astronomy*. The trigonometry-calculation connection was based on relations like

$$\sin A + \sin B = 2 \sin([A + B]/2) \cos([A - B]/2).$$

(See Exercise 1. See also Boyer for details of the calculation method.)

b) multiple-angle formulas

Viète created formulas for cosine and sine of multiples of an angle, in a way that combined combinations and trigonometry.

To understand the formulas, we need to look back at Pascal’s triangle (section IV.B.3).

Denote entry #*r* in row #*n* by C_r^n. (Remember that the top row is row #0, and the leftmost entry in any row is entry #0.) We said back there that the pattern—the recursion—in the triangle is that each entry is the sum of those just above it, left and right. In this new notation, the recursion is

$$C_r^n = C_{r-1}^{n-1} + C_r^{n-1} \qquad \text{if } r \text{ is from 1 to } n-1 \text{, with} \qquad C_0^n = C_n^n = 1.$$

(The C_r^n notation fits the arrangement of the entries in the triangle as binomial coefficients. However, it actually owes to the connection between them and **combinations** (subsets). We will not explore the connection.)

Write the familiar double-angle formulas

$$\cos 2\theta = \cos^2\theta - \sin^2\theta \qquad\qquad \sin 2\theta = 2\sin\theta\cos\theta.$$

The cosine formula has (forgetting the minus sign) coefficients 1, 1. The sine has coefficient 2. The three coefficients 1, 2, 1 form the #2 row in the triangle. Hence we may write

$$\cos 2\theta = C_0^2\cos^2\theta - C_2^2\sin^2\theta, \qquad\qquad \sin 2\theta = C_1^2\sin\theta\cos\theta.$$

Now apply the sum formulas to get

$$\begin{aligned}\cos 3\theta &= \cos\theta\,(C_0^2\cos^2\theta - C_2^2\sin^2\theta) - \\ &\qquad \sin\theta\,(C_1^2\sin\theta\cos\theta)\\ &= C_0^2\cos^3\theta - (C_2^2 + C_1^2)\cos\theta\sin^2\theta\\ &= C_0^3\cos^3\theta - C_2^3\cos\theta\sin^2\theta\end{aligned}$$

$$\begin{aligned}\sin 3\theta &= \sin\theta\,(C_0^2\cos^2\theta - C_2^2\sin^2\theta) + \\ &\qquad \cos\theta\, C_1^2\sin\theta\cos\theta\\ &= (C_0^2 + C_1^2)\sin\theta\cos^2\theta - C_2^2\sin^3\theta\\ &= C_1^3\sin\theta\cos^2\theta - C_3^3\sin^3\theta.\end{aligned}$$

(Do Exercise 3 to reconcile the cos 3θ formula with the one from VI.C.4a.)

You can see the pattern beginning to emerge. The pairs of formulas have an odd resemblance to a separation of the binomial expansion

$$(\cos\theta + \sin\theta)^n = C_0^n\cos^n\theta + C_1^n\cos^{n-1}\theta\sin\theta + C_2^n\cos^{n-2}\theta\sin^2\theta + C_3^n\cos^{n-3}\theta\sin^3\theta + \ldots$$

into two sums, one adding the even-numbered terms, the other adding the odd-numbered, each sum then supplied with alternating signs. Thus, from term to term in each formula, the sign changes, the coefficient skips one place in the triangle (with the sine formula holding the entries missing from the cosine), and the powers of cos θ and sin θ decrease and increase, respectively, by 2.

We extend the pattern by using the sum formula for the next level. (Beyond that, see Exercise 4.)

$$\begin{aligned}\cos 4\theta &= \cos\theta(C_0^3\cos^3\theta - C_2^3\cos\theta\sin^2\theta) - \\ &\qquad \sin\theta(C_1^3\sin\theta\cos^2\theta - C_3^3\sin^3\theta)\\ &= C_0^3\cos^4\theta - (C_2^3 + C_1^3)\cos^2\theta\sin^2\theta + \\ &\qquad C_3^3\sin^4\theta\\ &= C_0^4\cos^4\theta - C_2^4\cos^2\theta\sin^2\theta + C_4^4\sin^4\theta\end{aligned}$$

$$\begin{aligned}\sin 4\theta &= \sin\theta(C_0^3\cos^3\theta - C_2^3\cos\theta\sin^2\theta) + \\ &\qquad \cos\theta(C_1^3\sin\theta\cos^2\theta - C_3^3\sin^3\theta)\\ &= (C_0^3 + C_1^3)\sin\theta\cos^3\theta - \\ &\qquad (C_2^3 + C_3^3)\sin^3\theta\cos\theta\\ &= C_1^4\sin\theta\cos^3\theta - C_3^4\sin^3\theta\cos\theta.\end{aligned}$$

Recall that we used a triple-angle formula to solve some cubic equations (section VI.C.4b) and a quadruple-angle formula to solve a quartic (Exercise VI.C.4:4). Read Boyer on how Viète adapted the (45θ)-formulas to answer a (publicly announced) challenge to solve an equation of degree 45.

Exercises VI.C.5

1. Prove the following relations. (Hint for both: $A = [A + B]/2 + [A - B]/2$.)
   a) $\sin A + \sin B = 2\sin([A + B]/2)\cos([A - B]/2)$.
   b) $\cos A + \cos B = 2\cos([A + B]/2)\cos([A - B]/2)$.

2. Prove that
   $$c\sin A + d\cos A = \sqrt{(c^2 + d^2)}\,\sin(A + \tan^{-1}[d/c]).$$
   (Hint: Pretend *c* and *d* are positive, and draw a right triangle with legs *c* and *d*. This formula comes up in description of electrical signals.)

3. Show that
$$C_0^3 \cos^3\theta - C_2^3 \cos\theta \sin^2\theta = 4\cos^3\theta - 3\cos\theta \qquad \text{(both equal to } \cos 3\theta\text{).}$$
4. Prove by induction that
$$\cos n\theta = C_0^n \cos^n\theta - C_2^n \cos^{n-2}\theta \sin^2\theta + C_4^n \cos^{n-4}\theta \sin^4\theta + \dots,$$
$$\sin n\theta = C_1^n \sin\theta \cos^{n-1}\theta - C_3^n \sin^3\theta \cos^{n-3}\theta + C_5^n \sin^5\theta \cos^{n-5}\theta + \dots.$$
Each sum stops before the exponents turn negative, at ($\pm\sin^n\theta$) or ($\pm n \cos\theta \sin^{n-1}\theta$).

Section VI.D. The Astronomers

It is coincidence that this part is ending like the first, with discussion of the work of astronomers. Aside from the nice symmetry, though, the work leads naturally into developments beyond the Renaissance. [I have never read a better account of what follows than the **Ferris** chapters on Copernicus, Kepler, and Galileo, pages 61-101.]

1. Copernicus

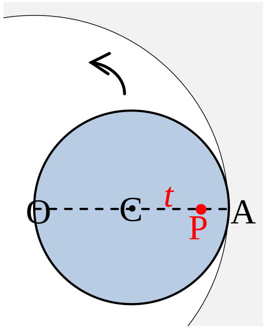
The phrase "paradigm shift" is overworked nowadays, used by those with things to sell to brag that their wares spring from revolutionary thinking. The classic paradigm shift is the introduction to Europe of the heliocentric model proposed by Copernicus (Mikolai Kopernik?, 1473-1543).

a) man and mathematics

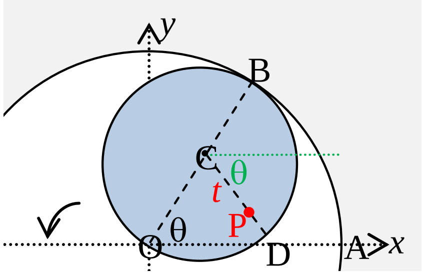
Recall that Aristarchus proposed heliocentrism around the time Archimedes was born. It was not accepted then—even Aristarchus reasoned in Earth-centered terms—but it is appropriate to link him with Copernicus. Each was an outstanding mathematician. Indeed, Copernicus was an extraordinary polymath: He was an expert in civil law, Church law, government finance, astronomy, medicine, and languages. Numerous leaders sought his advice. We have alluded to his trigonometry book. That book was actually part of *De Revolutionibus Orbium Celestium* (*Of the Revolutions of the Heavenly Spheres)*, which Copernicus hesitated to publish. It took his student Georg Joachim (Iserin) de Porris (1514-1576), the physician who took the place-name "Rheticus," to encourage him to publish the trigonometry part. Rheticus then produced in 1540 a kind of introduction to *Revolutions*, called *Narratio Prima* (*First Account*). (His *…De Triangulis* came later.) The success of the trig book and *Narratio* persuaded Copernicus to allow publication of *Revolutions*, which actually appeared in the year he died.

("Copernicus" is, like "Rheticus," a Latinized **toponym** [from the Greek roots for "place" and "name"]. Copernicus was born in a copper-rich region.)

To see some of his mathematics, consider the motion of a circle within a second circle twice as big, the interior circle rolling without slipping along the bigger. Copernicus derived a result attributed to the Persian astronomer Nasr al-Din al-Tusi (1201-1274): A point affixed to the interior of the smaller circle will trace out an ellipse as that circle rolls.



In the figure at left, we have a stationary circle (filled in white) with center O, rightmost point A, and radius OA = 2. A circle (shaded blue) of radius 1 starts out internally tangent to the bigger circle at A. Its center is C. Point P is attached to the blue disk, distance *t* from C, lying at first along OA. The smaller circle is going to rotate clockwise, so as to roll counterclockwise around the larger circle. In doing so, the smaller one causes C to trace out a circle and P some path through the interior of the larger circle.



At right, the blue circle has rolled around to put the point of tangency at B. Segment CP has rotated to the position shown, and we have extended it to meet the blue circle at D. Finally, we want to use our modern tools: We add coordinate axes with the origin at O.

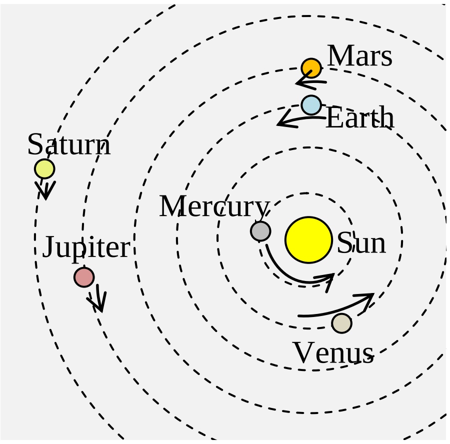
Radii OB of the big circle and CB of the small one must lie along the same line. (Why?) Let θ be the (radian) measure of angle BOA. Then C has arrived at coordinates

$$x = \cos\theta, \qquad y = \sin\theta.$$

This is not a surprise; we knew that C traces out the circle of radius 1 centered at O. What is important is that we can use those coordinates to locate P.

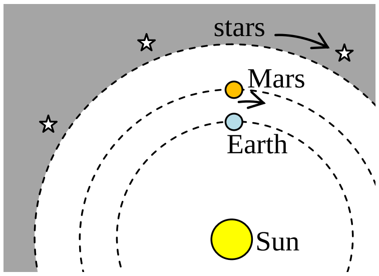
> In the larger circle,
>
> arc AB = 2θ.
>
> In the smaller,
>
> arc BD = 1(angle BCD).
>
> But arc AB and arc BD are equally long; they both equal the "rolling distance" of the inner circle. Therefore angle BCD = 2θ. Since the angle between CB and the (dotted green) horizontal at C must match θ (Why?), we conclude that the angle between CP and that same horizontal is likewise θ. That means P is (t cos θ) further right than C, (t sin θ) lower than C. The coordinates of P are therefore
>
> $$x = \cos\theta + t\cos\theta = (1+t)\cos\theta, \qquad y = \sin\theta - t\sin\theta = (1-t)\sin\theta.$$
>
> Observe that if t is 1—if P is actually on the small circle—then y is constantly 0. In that case, P just patrols the diameter from A to (-2, 0). (The figure shows D close to the x-axis. It actually has to be right on the axis.) In the expected situation, with $0 < t < 1$, the coordinates of P satisfy
>
> $$x^2/(1+t)^2 + y^2/(1-t)^2 = cos^2\theta + sin^2\theta = 1.$$
>
> The point traces out an ellipse.

b) the solar system model



Copernicus proposed that the planets travel along circular orbits surrounding the Sun, arrayed in the order shown in the figure at right. He had the orbital speeds (suggested by the lengths of the arrows) decreasing toward the outside: Mercury moving fastest, Venus second fastest, and so on. Such a model is consistent with—in fact, it explains—a number of our observations.

> 1. Venus in our sky is always close to the Sun. Mercury is even closer, so much so that Mercury is hard to spot in the twilight.
>
> 2. Mercury orbits so fast that it completes multiple observational cycles during one (Earth) year.
>
> To understand "observational cycles," note that in the figure, the Earth-Mercury line is tangent to the Mercury orbit. With that position, Mercury stands in our view as far rightward (west) from the Sun as it can seem. The situation is called "greatest westward elongation." Those elongations recur three or four times each year, on average about 116 days apart.
>
> Venus completes the same cycle in about 584 days (1.6 years). It completes more than one orbit in our year, but less than one observational cycle.
>
> 3. Even more important is that the model can explain the motion of the outer planets without resorting to the epicycles of Apollonius, Hipparchus, and Claudius Ptolemy (section III.C.4a).



In the next figure, we spin our view counterclockwise at the orbital rate of Earth. That rotating frame of reference makes Earth appear to be stationary at the top of its orbit. We have Mars (or Jupiter or Saturn) at what is called "opposition," when the planet and the Sun stand in opposite directions in our sky. Our rotating reference frame also causes the starry background seemingly to rotate clockwise, at Earth's rate: once around in a year. At the same time, Mars appears to revolve clockwise, because

revolution rate of planet – revolution rate of Earth

is a negative number; but the complete revolution takes more than a year, because that negative number has smaller absolute value than its second term.

Those rotation rates mean that on average, the *apparent clockwise angular speed* of Mars is lower than that of the stars. On the average, therefore, we see Mars drifting counterclockwise (eastward) *as viewed against the starry background*. But for a period (months for Mars, weeks for the others) surrounding opposition—when the planet is closest to Earth for that year—the planet is so close to Earth that its angular speed exceeds the angular speed of the stars. Consequently during that period, the planet advances clockwise (westward) relative to the stars. Explaining that westward ("retrograde") drift was the whole point of the epicycle model.

However, Copernicus could not do away with the epicycles entirely. There are mismatches between the circular model and observation. For one, if the planets actually revolved in circular orbits centered at the Sun, then at every opposition of a given planet, it would stand at the same distance from Earth. The distance would be the difference between the orbital radii. Assuming the planet does not have variable reflectivity, it would have to appear equally bright at each opposition. For Jupiter and Saturn, the opposition brightness is reasonably constant. For Mars, the variation is unmistakable. At opposition in 2003, Mars was more than twice as bright as Sirius, the brightest (nighttime) star. In 2010-12, it was about 83% as bright as Sirius. For the 2016-18-20 oppositions, it will considerably outshine Sirius again. (Those will not reach 2003. That was a historically brilliant apparition; see NASA's article. But the 15-year cycle is real.) In this phenomenon, the Copernican model was in conflict with the God of War.

You could reconcile the circles and the evidently variable opposition distance via a modification: Give the circles unequal centers. Unfortunately, that begins to complicate the model, and still does not fully account for the observed positions of the planets. Copernicus could not completely abandon the epicycles (along circles surrounding the Sun instead of surrounding Earth).

2. Kepler

During the last quarter of the sixteenth century, an eccentric Dane named Tycho Brahe established and ran an observatory at Uraniborg (in what is now Sweden). He compiled a fantastic record of precise astronomical measurements, all by naked-eye observation. Based on those, he proposed a solar system ruled by an odd hybrid of Copernican and Ptolemaic elements.

A German named Johannes Kepler (1571-1630) wanted to analyze the record. He believed the Copernican model, not Tycho's hybrid, and figured the data would validate Copernicus. He became a sort of apprentice to Tycho after the latter moved to Prague. (Tycho treated him more like a slave.) At Tycho's death, Kepler finally got his hands on the planetary data. He set out to reconcile them with the Copernican model.

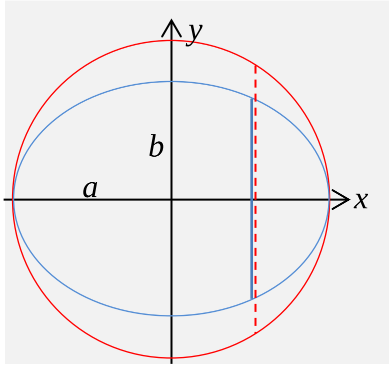
They would not fit. For years, he tried to squeeze Tycho's observations into circular orbits, without success. The positions were more suggestive of ellipses.

a) man and mathematics

Kepler was qualified to know a conic section when he saw one. He applied to conics a marvelous intuition with the infinite and the infinitesimal; see **Boyer** 354-357.

> Recall our treatment of the conic sections (section III.A.7). Kepler thought of the horizontal section, the circle, as the extreme case in which the two foci coincide. Tilting the cutting plane separates the foci, to create the ellipse. Further tilting moves one focus farther out, until the inclination of the plane matches that of the element of the cone; at that stage, the remote focus reaches infinity. Continued tilting makes the remote focus come back from infinity, *in the half of the cone opposite the near focus*. Finally, shifting the plane so that it crosses the vertex of the cone produces the other extreme case, two intersecting straight lines (at whose intersection the two foci again coincide).

He thought in infinitesimal terms to measure the area of the ellipse.



> At right we have a circle (red) of radius a and an ellipse (blue) with semiaxes $a \geq b$, both centered where we now put the origin of a coordinate system. Kepler thought of the circle's area as composed of an infinity of **infinitesimals**, the (dashed red) vertical chords. Similarly, the area of the ellipse is composed of the blue chords.
>
> In our terms, the circle and ellipse have equations, respectively,
>
> $x^2 + y^2 = a^2$ and $x^2/a^2 + y^2/b^2 = 1$.
>
> Therefore the vertical chords at a given x-value have lengths
>
> $2y = 2\sqrt{}(a^2 - x^2)$ and
>
> $2y = 2\sqrt{}(b^2[1 - x^2/a^2]) = 2\, b/a\, \sqrt{}(a^2 - x^2)$.
>
> The height of each constituent chord of the ellipse is b/a times the height of the corresponding chord of the circle. Hence the area of the ellipse is b/a times the area of the circle. The area of the ellipse is
>
> $A = b/a\, (\pi a^2) = \pi ab$.

b) the planetary laws

Failing to match the orbit of Mars with a circle, Kepler decided to try to fit an ellipse to the orbit. The match was nearly perfect. By about 1602, he discerned the first two of the laws that govern the motion of the planets. He published those two in 1609.

Kepler's First Law: **Each planet orbits the Sun along an ellipse …**

That is half the First Law. Symmetry-minded as we are, we might expect the Sun to be at the center of the ellipses. The second half of the Law says:

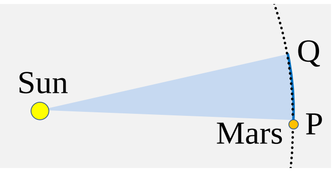
… with the Sun at one of the foci.

> That principle takes care of the varying opposition brightness. We now know that Earth's orbit is nearly circular, with an eccentricity of 0.017. For Mars, the eccentricity is 0.093. *That is still nearly circular*; see Exercise III.A.7: 5b. What is more salient with the greater eccentricity is the offset of the Sun from the center of the ellipse. Earth's distance from the Sun varies roughly from 91.4M (when Earth is at "perihelion," the orbit's major vertex closer to the Sun) to 94.5M miles ("aphelion," the remote major vertex); see Exercise 2. The corresponding Mars distances are 128M and 155M. Those numbers imply that when Earth brings Mars to opposition under the latter's perihelion, the distance between us and Mars is around 128M – 93M = 35M miles. With Mars at aphelion, the distance is about 155M – 93M = 62M. If Mars is equally reflective of the Sun's light at both stations, then the brilliance of Mars to our eyes drops to $(35/62)^2 \approx 32\%$ from a close opposition (like 2003) to a distant one (2010).

Having dispensed with circular paths, Kepler then did away with uniform speeds.

Kepler's Second Law: Each planet moves so that the focal radius from it to the Sun sweeps out equal areas in equal times.

In the figure at right, we have a planet, say Mars (orange), tracing the arc PQ (solid blue) along its orbit (dotted), in a time t. In that time, the focal radius sweeps out the blue-shaded region of area A. The Second Law states that the sweep rate A/t is constant throughout the orbit. It dictates the relative speed of each planet at different points of its orbit.



> If t is small, then arc PQ is indistinguishable from its chord. (For example, if t is one hour, then the Mars arc is a mere 54000 mi (average) long. That means it spans an angle
>
> 54000/(140 million average) $\approx$ 0.02°.)
>
> Moreover, if P is at either of the major vertices of the ellipse, then the region is indistinguishable from a right triangle. At those places, the tangent to the ellipse (and therefore the direction of PQ) is perpendicular to the major axis.
>
> At aphelion, we said, the focal radius for Mars is about 155M mi. Denote the planet's speed PQ/t by v_a. Then the sweep rate is
>
> (area of right triangle)/t = (1/2 PQ 155M)/t = 1/2 (155M) v_a.
>
> At perihelion, the focal radius is about 128M. Call the speed there v_p. The sweep rate there is
>
> (1/2 PQ 128M)/t = 1/2 (128M) v_p.
>
> Setting the rates equal, we have
>
> v_a / v_p = 128M/155M.
>
> At those places, the planet's speed is inversely proportional to distance from the Sun.

Kepler first thought that the last relation applies at all points of the orbit. It was later that he realized that the relation fails at the other points, where the planet's direction is not perpendicular to the focal radius. It is area swept per unit time that remains constant, and not the product of speed and distance.

The Third Law unifies the system, fixing the speed of each planet *in comparison with the others*.

Kepler's Third Law: The square of the orbital period of a planet is proportional to the cube of its major axis.

> More simply[?], the period is proportional to the 3/2 power of the major axis. Thus, Earth's orbit has major axis 186M, Mars 283M. Therefore the period of Mars is $(283/186)^{3/2} \approx$ 1.88 times the period of Earth. In 1.88 years, there are roughly 687 days.

The Third Law had a momentous consequence. It showed us the *scale* of the solar system. The period of a planet is something we can *observe*, by plotting it against the stars. From the periods, we deduce the relative sizes of their orbits, compared to Earth's (Exercise 4).

Exercises VI.D.2

1. At its elongations, Venus (seen from Earth) stands between 45° and 48° from the Sun. Mercury, with a more eccentric orbit, varies from 18° to 28°. (Copernicus made his own accurate observations. He could have obtained measurements like these.) Assume that each planet's greater number goes with its aphelion, smaller number with perihelion, and pretend that Earth's distance from the Sun is always 93M mi.
   a) Find the length of each planet's major axis.(Hint: Remember that at elongation, the Earth-planet line is tangent to the planet's orbit, and that aphelion and perihelion are the major vertices of the orbits.)
   b) Find each planet's orbital period. (Modern values for them are 224.7 and 88.0 days.)
   c) Find their orbits' eccentricities. (Modern: 0.007 and 0.205)

2. Given that the orbital eccentricities of Earth and Mars are 0.017 and 0.094 and the major axes 186M and 283M miles respectively:
   a) Show that the distances from the Sun to the two major vertices of Earth's orbit are roughly 91.4M and 94.5M miles.
   b) Show that the corresponding Mars distances are 128M and 155M miles.
3. Use the data given in Exercise 2.
   a) For Earth, perihelion comes each revolution around January 4. How much more **insolation** (solar energy) does Earth receive then, compared to what it receives around aphelion near July 4? (Insolation rate drops with the square of distance from the Sun.)
   b) By coincidence, the perihelion of Mars happens during the northern hemisphere's winter, just as on Earth. How much more insolation does the planet get at perihelion than it gets at aphelion?
   c) Assume that the (linear) speeds of Mars at perihelion and aphelion are 1.45M/day and 1.2M mi/day. What are its corresponding angular speeds (degrees per day) around the Sun? (These speeds and the insolation rates from (b) imply that for the southern hemisphere of Mars, winter is bad news; it is both colder and longer than for the northern.)
4. a) Jupiter takes about 12 years to orbit the Sun. How long is the major axis of its orbit?
   b) Saturn takes about 29 years. How long is the major axis of Saturn?

3. Galileo

a) man and mechanics

Galileo (1564-1642) became professor of mathematics at Pisa. There, he angered the faculty by his open disdain for ideas of Aristotle, which of course prevailed among his colleagues. Shown the door by those colleagues, he gladly accepted appointment at Padua. That was near Venice, and therefore a step up in prestige. Hunger for recognition and fame defined Galileo's life, and substandard deference to the authority of authorities nearly ended it.

(i) falling objects

One time at chapel during his short-lived medical training, Galileo watched the chandeliers swaying with the occasional breeze. He decided to time their swings by counting his pulse. He made a striking observation: As long as a chandelier did not swing *too* far, the **period** of its oscillation was independent of the **amplitude**. In shorter words, the time it took to swing back and forth was the same, no matter how far it swung.

The observation led Galileo to experiment with pendulums, something you can try yourself. Tie a heavy "bob," like a bolt or nut, to a length of string. Hold the other end of the string with your fingers, and brace that hand. (The ideal is that the thickness of the string and dimensions of the bob should be insignificant compared to the length of the string, the bob much heavier than the string, and the braced support stationary.) Then displace the bob an inch or so, and count how many swings the bob makes in 30 seconds. Do the same with an initial displacement of four inches or so. You should find roughly the same number of oscillations for the two displacements. It is easy to see that the pendulum makes the longer swings with greater speed. It turns out that the greater speed exactly offsets the bigger distance.

> Separately, make the hanging string either a quarter or four times as long. It is obvious that the longer string creates a higher period. Counting the oscillations will give you a good idea of the quantitative relation between period and length, a relation that must have been known since ancient times.

Then Galileo made a more important observation. Measure the hanging string and count again the oscillations. Then replace the bob with a significantly heavier one, or with two bobs, and match the length of the string. Your count should point to the same period. You get the like result by switching instead to a lighter bob. For a given length of string, Galileo discovered, the pendulum's period is independent of the bob’s weight.

Galileo realized that the pendulum bob *is falling*, its fall arrested by the string. From further experimentation with falling objects—none involving the Tower of Pisa—he drew a conclusion that contradicted Aristotle:

Except for the effect of air resistance, light objects and heavy objects fall at the same rate.

(About that air-resistance proviso: Tear one third from an ordinary sheet of paper. Crumple the smaller piece into a ball. Then drop the flat two-thirds and the crumpled piece from the same height. You can see that the air slows the heavier piece’s fall. Alternatively and in modern terms, consider an airplane dropping food packages on a relief mission. The crew slows the descent of the packages by adding to them the weight of parachutes.)

Galileo argued for this principle by means of **thought experiments**, which we now associate more with Einstein. Imagine a heavy ball dropped from some height. It will fall at the rate of a heavy object. Now imagine cutting the ball into halves, then connecting the two halves by means of a small string. The halves will fall together at the rate of light objects. But surely the halves are not conscious of no longer constituting a single ball. The ball-string-ball combination must therefore fall as though it had remained the one heavy ball. It must be that the rate of fall for one half ball equals the rate for the doubly heavy whole ball.

[Centuries later, David Letterman added evidence. He was videotaped on a rooftop dropping a six-pack of light beer together with a six-pack of regular beer. The simultaneous splats demonstrated the equal falling rates.]

(ii) speed of fall

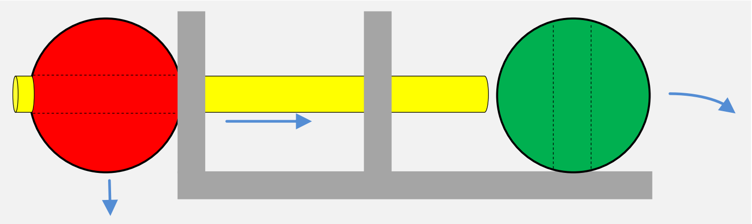
Galileo then experimented to investigate actual speed of fall. Our own experience tells us that objects fall way too fast to allow timing with anything but electronics. Galileo slowed the fall by letting objects roll down inclined planes. Naturally, steeper inclines led to faster falls, but Galileo noticed a consistent pattern: Whatever distance the object rolled in the first time interval, it rolled three times as much in the second (equal) time interval, five times as much next interval, …. You can instead look at it in terms of the total distances covered,

$$1, \qquad 1 + 3 = 4, \qquad 1 + 3 + 5 = 9, \quad \ldots.$$

Thus, speed increases with time at constant rate, distance covered increases as the square of the time. Recall that Oresme (section V.B.3c) had described that connection. Galileo made no reference to Oresme, but he must have been aware of the latter’s work.

(iii) independence of components

Galileo had a knack for tinkering. His designs for experimental apparatus were brilliant, and none more so than a gizmo that is a staple of our school physics labs. (See it advertised at UniScience Laboratories.) The device has a spring-loaded rod, shown yellow and in the cocked position in the figure at left. In that position, its left end holds up a ball (shown red) through a hole in the ball. Its right end adjoins a similar ball (green). When the rod is triggered, it pops rightward,



simultaneously dropping the red ball vertically and propelling the green one horizontally to the right. If the device and the floor are both level, then the falling red ball and the flying green ball hit the floor at the same time, irrespective of the height of the device. In addition, the horizontal distance covered by the green ball is proportional to the square root of the device's height.

Multiple conclusions follow. First, the coincident impacts say that the falling rate of the green ball is unaffected by the fact that it is moving horizontally as well. It next follows, in view of Galileo's discovery that the red ball's falling distance is as the square of the time, that the green's horizontal distance is a multiple of the time. That multiple is evidently the initial horizontal speed times the time. Therefore the horizontal motion is unaffected by the vertical falling. Thus, in the language of physics, the green ball's **vertical and horizontal components of velocity** are independent.

> Galileo proceeded to describe the path of the green ball. If v_0 is the imparted horizontal speed, then the horizontal distance covered in time t is v_0t. The vertical speed increases at a constant rate, the **acceleration** a (which Galileo could not conveniently measure). After time t, the acquired vertical speed is at. By Oresme's argument, the average speed is half that, $at/2$. Therefore the distance fallen is $(at/2)t$. If we establish our kind of coordinate system at the top of the flight, then we locate the green ball at coordinates
>
> $$x = v_0t, \qquad y = at^2/2.$$
>
> Accordingly, the path of the flying ball is the parabola given by
>
> $$y = a(x/v_0)^2/2 = -a/(2v_0^2)\ x^2.$$

Arguing from symmetry, Galileo predicted that any object in free fall—an object subject during flight only to the influence of its weight—traces a predictable parabolic trajectory (Exercise 2). Always alert for opportunities to turn his ideas into money, he produced tables and devices by which an artillery crew could calculate the muzzle velocity of its cannon, then calculate the angle at which to elevate the gun to hit (or as the Pentagon says, "service") a target at given distance.

b) the solar system

In 1609, Galileo started to hear reports of telescopes. They were an Arabic invention, conveyed to Europe by the Dutch (section V.A.4a). Always alert for opportunities …, Galileo looked into their design, built his own models, and began to sell them as spotting instruments for commercial and military lookouts. He gave some to influential people who might further his sales and career. Then he turned his telescope to the night sky.

There he *saw* contradictions to Aristotle. For one, the Moon was not a perfect sphere. It was rough, strewn with craters and jagged peaks. Behind it, there were stars revealed by the telescope that were invisible to the naked eye. Galileo took them as evidence that the stars are not fixed to one sphere, but are instead scattered throughout depths of space. (Tycho had also spotted evidence against Aristotle's model of the universe; see **Ferris**, pages 69-73.)

In 1610, he aimed his telescope at the planets. There, he found support for Copernicus. Looking at Jupiter, he found that it had star-like companions. They visibly shifted position over the course of hours; they clearly were not stars. Observing them night by night, he realized that the **satellites** (Kepler's later word for them, from the Latin for "attendant" or "follower") dance *around* Jupiter. He had found the first evidence of a system of celestial objects in orbits centered at another one, not at Earth. The system was a miniature of the Copernican solar system. Later that year, Galileo looked at Venus. He detected the planet's phases. He read them correctly: Venus orbits the Sun. When Venus is beyond the Sun, we can see most of its sunlit face and it looks nearly "full". At its elongations, we see half the illuminated face, and it looks like our quarter moon. When it comes around to our side of the Sun, we see a fraction of its lit face, and it appears as a crescent.

[Galileo (remember “Always alert …”) came up with a way to apply his astronomy to the problem of determining longitude. The moons of Jupiter sometimes hide behind the planet, and sometimes they disappear in space by entering its shadow. When a moon slides behind the right-hand (westerly) edge of Jupiter as seen from Prague, it is still visible by parallax from places further right, like Paris. Not so with an **eclipse**; when a moon enters the shadow, its disappearance is at once visible (ignore light speed) from every place on Earth whose view is not blocked. Galileo figured that he could produce an accurate table of eclipse times, say for Venice. Suppose an observer saw a disappearance two hours earlier, according to his own clock, than the predicted time. He could judge that he was two hours west of Venice, in the Atlantic at a longitude 2/24 × 360° = 30° to the west. The idea was scientifically unimpeachable, practically worthless. Galileo did not describe how to overcome the problem of trying to train a telescope from the rocking deck of a ship, nor how to keep accurate time at sea. The latter problem proved to be the key. You should read the story of its solution in *Longitude: The True Story…*, by Dava Sobel. Her other books include *Galileo’s Daughter*.]

c) deference to authority

Galileo published his discoveries in *Siderius Nuncius* (*The Starry Messenger*). He then launched a campaign to spread the word that Copernicus was inarguably correct, and that the Roman Church had to admit its error and renounce its Scripture-based picture of the sky. In 1616, the Church responded by putting *Of the Revolutions …* on the Index of Forbidden Books. Scientific minds did not praise Galileo for persistence; they decried his stubbornness. He had managed to render largely unavailable a classic book that had circulated freely for seventy years.

Galileo pressed on in his insistence on dismissing Ptolemy. By 1632, he had published the *Dialogue Concerning the Two Chief World Systems*. In the book, he put defense of the Copernican system into the arguments of a reasonable, learned man, and defense of the Ptolemaic (and Church) system into those of a jackass named “Simplicio.” (It was especially dangerous that the book was in Italian, and therefore not limited to those educated in Latin.)

The Church had seen enough. It tried him for heresy and ordered him to recant his false theories or face burning at the stake. In 1634, he “abjured [his] heresies.” He spent his last eight years under house arrest. From his house, he published *Dialogue Concerning Two New Sciences*, which included the mechanics principles he had discovered in the 1590’s. It appeared after he went blind, probably owing to his dreadful habit of putting his telescopes directly on the Sun. (His discovery of sunspots had been further argument against the celestial perfection postulated by the Aristotelians.)

[It took until 1980 for the Catholic Church to announce that it had erred in persecuting Galileo. By contrast, *Of the Revolutions …* came off the Index in the 1750’s.

I must have read the following somewhere, but cannot remember the source. Investigation into heresy is made by the Vatican’s guardians of doctrine, now generally called the Holy Office (which sounds gentler than “Roman Inquisition”). Its head is an important Cardinal; the 2005 incumbent, Joseph Ratzinger, became Pope Benedict XVI. When *Of the Revolutions …* appeared, the question of its suitability was put to the Office. The leader had an interesting reaction. To declare that the Earth moves is plainly contrary to Scripture, he said, and therefore heretical. But to *pretend* that the planets circle the Sun, for the sake of explanation—to use heliocentrism as a hypothetical *teaching device*—is not merely acceptable, it is felicitous. It argues to the elegant, powerful simplicity of the Creator’s design.]

Exercises VI.D.3

1. The Church would have put Galileo to the fire for espousing heliocentrism, but Kepler’s Laws were twenty years old by then. Why did the Church not threaten to smoke Kepler?

2. Suppose a cannon discharges its ball with a speed V feet per second. Imagine it fired at an angle of elevation θ above level ground.
   a) Show that the projectile has an initial horizontal speed of $V \cos \theta$ and initial vertical speed $V \sin \theta$. (Use geometric reasoning, not "components.")
   b) Assume that the horizontal speed remains constant. Assume further that the vertical speed decreases by 32 ft/sec every second, and has an average during any time span equal to its value at the midpoint of the span. Find an equation for the trajectory. (Choose where to put your coordinate system.)
   c) Under the assumptions of (b), show that the ball takes the same time going up to its high point as it does descending to the ground.
   d) How far does the ball travel horizontally before landing?
   e) What is the range of the cannon (the greatest horizontal distance its ball can go)? (This is important to an artilleryman. It is how close he must allow the enemy to come, or how close he must approach, for his cannonballs to reach the enemy before bouncing. [They are useful to our man after they bounce, too].)

Section VI.E. Practical Men, Contemplative Men

Merzbach has some wonderful lines about mathematical progress. Their context is the development of geometry, but you can see their enduring relevance. Beginning on page 8, Dr. Merzbach writes:

> "The debate, extending well beyond the confines of Egypt, about whether to credit progress in mathematics to the practical men (the surveyors, or "rope-stretchers") or to the contemplative elements of society (the priests and philosophers) has continued to our times. As we shall see, the history of mathematics displays a constant interplay between these two types of contributors."

Viète and Galileo span both groups. Judge for yourself where in the two groups you might put any of the contributors we have met, from Brahmagupta through the astronomers.

Chapter VII. The Road to Calculus

We have reached 1600. Europe begins to look the way we think of it now. The national states have mostly formed. Spain is the most powerful of them. Its New World empire covers half of South America, all of Central America, and the North American expanse from Mexico into Oregon, plus Florida. In 1585, Spain had held much of Italy, along with Holland and the Philippines. Three years later, it sent a giant armada to conquer England. The failure of that venture was the dawn of England's dominion of the seas, and inspired the Dutch to break free. Both England and Netherlands then set about making their own conquests in the Americas and Asia. France had come together under a strong central government; the Bourbon line began in 1589, and the three-quarter-century reign of Louis XIV would start in forty years.

The state of mathematics was mixed. The extant geometry would have been recognizable to the Greeks. Algebra, however, had advanced into a theory of equations much different from what al-Khwarizmi studied (or would have considered useful). The most important discoveries in number theory were Greek, Indian, and Chinese, all of them transmitted to Europe through the world of Islam.

Still, by 1600 Europe had become undisputed leader in scientific and mathematical inquiry. The three mathematicians we called "astronomers" in Section VI.D were more than examples. They introduced a revolution in scientific thought. The revolution extended to the relation between math and science, as the latter began to drive mathematical progress.

Section VII.A. Areas and Tangents

In the seventeenth century, the ancient geometric studies of quadratures and tangents accelerated. They evolved from geometric to algebraic approaches. The geometers were willing to make intuitive, heuristic arguments, especially in using infinitesimals. The algebraists made more precise, rigorous arguments. Interestingly, **Struik** (pages 100-101) describes both philosophies as reflective of Archimedes. The rigor was characteristic of the writings of Archimedes, as in the argument in section III.A.6c(ii) that the volume of the cone is greater than any number below (1/3 × height × base) and smaller than any number above. The intuition was what the Syracusan, in *The Method*, espoused as the mother of discovery. The distinction is striking, because Latin versions of the writings had existed for centuries, but *The Method* had disappeared and was not rediscovered until around 1900.

1. Cavalieri

Two geometers were students of Galileo. The first was Buonaventura Cavalieri (1598-1647).

Cavalieri studied with Galileo, but adopted the methods of Kepler. His influential *Geometria Indivisibilibus Continuorum* (*Geometry* [*via*] *Indivisibles of* [*Continuous Objects*]) appeared in 1635, when he was professor at Bologna. He was as willing to add up infinitesimals ("indivisibles") as Kepler had been. Recall Kepler's argument (section VI.D.2a) for the area of the ellipse. Extending another argument of Kepler—picturing a circle as the sum of an infinity of infinitesimal triangles with common apex at the center—Cavalieri viewed the sphere as the sum of an infinity of pyramids or cones having apex at the center. For each such solid, the height matches the radius of the sphere. Each has volume

1/3 × height × base = 1/3 × radius × base.

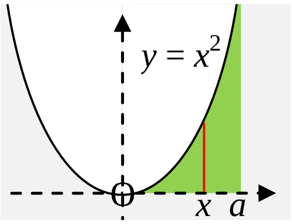
Summing those indivisibles, Cavalieri said the volume of the sphere is

1/3 × radius × (sum of bases) = 1/3 × radius × (area of sphere).

(Check that against our formulas for the volume and area. Notice also the resemblance to the relation

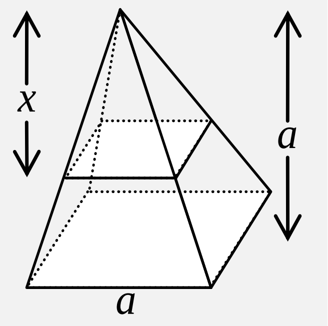
area of circle = 1/2 × radius × circumference

from section III.A.4a.)



To see his indivisibles in action, view the figure at right. It shows the region (green) under the graph of $y = x^2$ between $x = 0$ and $x = a$. For Cavalieri, the region was composed of an infinity of vertical indivisibles, the one (red) at x having length x^2. To add those (x^2)'s, he drew the square-based pyramid in the figure below. At the level x below the apex, x^2 is the area of the square cross-section. Therefore the sum of the lengths x^2 is the same as the sum of the like-valued areas. Those areas are the indivisibles that constitute the volume of the pyramid. Hence the area under the graph is the volume of the pyramid,

$$1/3\ a\ (a^2) = a^3/3.$$



[**Boyer** calls Cavalieri's indivisibles "quasi-atomic" entities. **Struik** (page 48) notes that *The Method*, which was a letter from Archimedes to Eratosthenes, has been construed as opposing a school associated with Eudoxus and Democritus. It was Democritus, of course, who proposed the existence of atoms.]

Notice that thinking of the solid as the sum of its horizontal cross-sections leads immediately to what we now call **Cavalieri's principle**: If two solids have the same height (vertical extent), and have at each horizontal level cross-sections of equal areas, then the two solids have equal volumes.

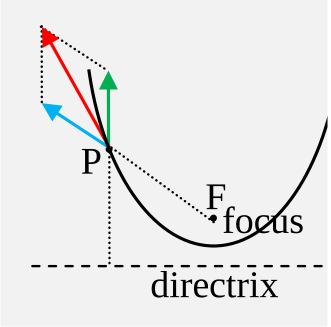
Cavalieri also squared other power-curve regions, and eventually drew the conclusion that the area under the graph of $y = x^n$, from $x = 0$ to $x = a$, is $a^{(n+1)}/(n + 1)$. Read Boyer 363-4 to see a generalization of Cavalieri's method by which he squared the spiral of Archimedes.

2. Torricelli

Evangelista Torricelli (1608-1647) met Galileo toward the end of the old man's life. Torricelli had been impressed by the *Dialogue Concerning Two New Sciences* (section VI.D.3c), which sparked his interest in the sciences. It is there that his name is better known. He explained, measured, and used air pressure. He put a hollow column into a bath of mercury and exhausted the air in the column, above the mercury. He found that the pressure of the atmosphere, on the mercury in the bath, would push up into the column about 30 inches of mercury. That was the first barometer. With it he discovered, among other things, that atmospheric pressure varies from day to day. Those 30 inches are about 760 mm. A pressure 1/760 of that—enough to support 1 mm of mercury—is called a **torr**.

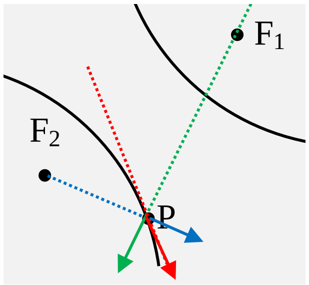
a) tangents

Torricelli described the tangent to the parabola in a way that harked back to Archimedes for the spiral (section III.A.6c(i)) and to Galileo for trajectories (section VI.D.3a(iii)). Recall that a parabola is a locus whose points are equidistant from a focus and a directrix. Therefore a particle P moving along the parabola (black) at right moves away from the focus F with the same speed as it moves vertically away from the directrix. As physics puts it, the vertical component (green) of the velocity is as long as the component (blue) in the direction FP. The combined motion ("resultant" velocity) is along the diagonal (red) of the parallelogram determined by the components. Since the parallelogram is a rhombus, the diagonal bisects the angle between the components. That means the direction of motion, and therefore the tangent, bisects the angle between the upward vertical and the extension of FP.



Notice that the tangent's orientation explains the parabola's reflection property. A ray striking the parabola at P (from the side that has the focus) gets reflected along a line making the same angle with the tangent as the ray. [Never mind that the Law of Reflection always refers instead to the *normal* and the complements of those angles.] Therefore if a ray comes to P along FP, then it is reflected vertically up, which means parallel to the parabola's axis. Similarly in reverse, if a ray comes down to P parallel to the axis, then it gets reflected toward F.

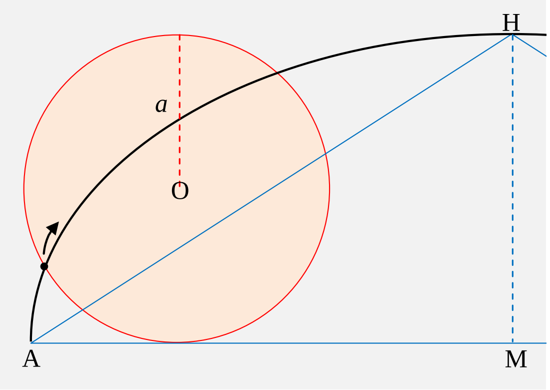
A like argument describes the tangent to the hyperbola and explains its reflection property. For the tangent, start with the figure at left. The black curves are the two branches of a hyperbola whose foci are F_1 and F_2. What characterizes the points P of the hyperbola is that the difference between the focal radii, $F_1P - F_2P$ in the figure, is constant. Accordingly, if P is moving along the lower branch as shown, then for every inch it adds per unit time to F_1P, it must add a like inch to F_2P. In other words, its component of velocity away from F_1 (green arrow) must be as long as the component away from F_2 (blue). Therefore the tangent (red) must bisect the angle between the focal radii. (Apollonius knew that.)



For the reflection property, think of the hyperbola as a mirror. Then a ray headed for one focus reflects so as to head for the other; and a ray originating at one focus reflects off the hyperbola so as to appear to be coming from the other. To see it in the figure, extend F_2P beyond P. Imagine a ray of light incoming along the extension *toward* P (opposite of the blue arrow). Because the tangent bisects angle F_1PF_2, the incoming ray and PF_1 make equal angles with the tangent. Such a ray, which unimpeded would continue to F_2, is reflected at P toward F_1. From the opposite side of the hyperbola, a ray originating at F_2 and hitting P gets reflected along the extension (green arrow) of F_1P, so as to appear to be coming from F_1.

b) area

Torricelli used indivisibles to square the cycloid. A **cycloid** is the path followed by a point pinned to a circle that is rolling along a baseline. As the circle rolls, the point traces a series of arches, symmetric about their midlines, with cusps at either end. [See a nice animation of it at Wikipedia®.] In the figure at right, the fixed baseline is AM. Initially, the circle sits with the pinned point (black dot) at position A. As the circle rolls to the right, the dot goes right and up (arrow), tracing out the black curve. After half a rotation, the dot arrives at its high point, position H. There half the arch ends. We may relocate M to put it vertically under H. The dot continues right and down, to complete the arch. We want the area of the arch.



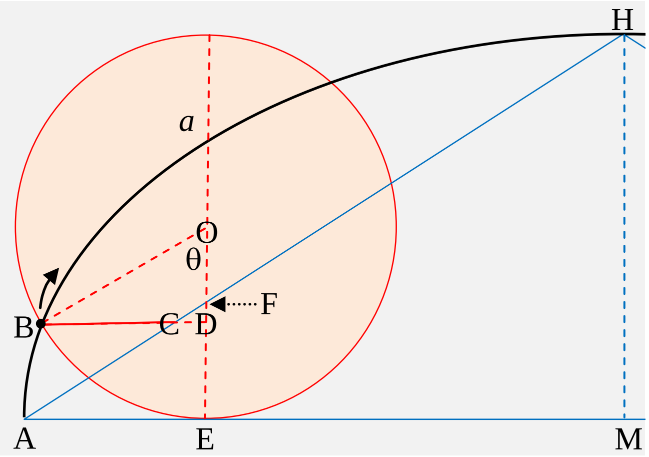
Part of the area is the **inscribed triangle** (outlined in blue), whose vertices are H, A, and the other end of the arch. Its height HM is clearly the diameter $2a$ of the circle. Its base is what the circle rolls out over a complete turn, namely the circumference $2\pi a$. Therefore its area is

$$1/2 \times 2a \times 2\pi a,$$

twice the area of the circle.

For the remaining area, outside the triangle, Torricelli turned to indivisibles.

Imagine in the previous figure that the circle has rotated through an acute angle of θ *radians*. That puts the circle in the position shown at right, with the dot at B, the point of tangency at E, and the angle BOE (O being the moving center) equal to θ.



Draw the vertical (diameter) at E and the horizontal at B. EO crosses the horizontal at D, at height

$$\mathrm{ED} \;=\; \mathrm{EO} - \mathrm{DO} \;=\; a - a\cos\theta.$$

EO crosses side AH of the inscribed triangle at F, at a height EF given by similar triangles:

$$\mathrm{EF}/\mathrm{AE} \;=\; \mathrm{MH}/\mathrm{AM} \;=\; 2a/\pi a.$$

Because AE is the distance rolled by the circle, it matches arc BE. Therefore

$$\begin{aligned}\mathrm{EF} \;&=\; (2/\pi)\,\mathrm{AE}\\ &=\; (2/\pi) \times \text{length of arc} \;=\; (2/\pi)\,a\theta.\end{aligned}$$

That implies that EF exceeds ED, and the picture is right.

[Draw for yourself, on one set of axes, the graphs of

$$y = (2/\pi)x \qquad\qquad \text{and} \qquad\qquad y = 1 - \cos x.$$

Use them to see that

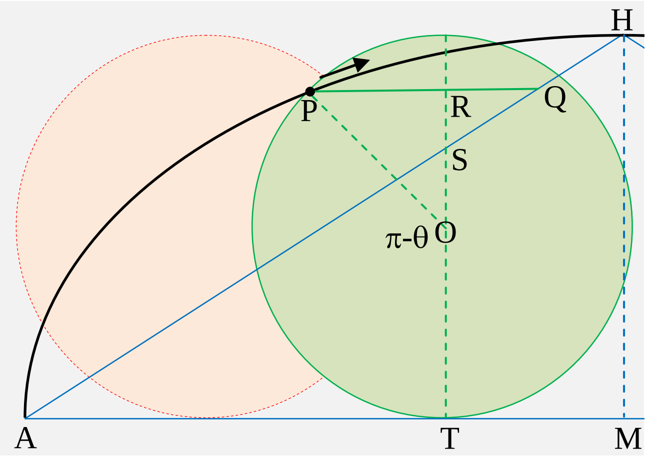
$$(2/\pi)ax \;>\; a - a\cos x$$

for x between 0 and $\pi/2$, and the opposite for x between $\pi/2$ and π. The "opposite" part will be important below.]

Having verified the figure, let C be where BD crosses side AH of the inscribed triangle. BC is an infinitesimal of (half) the area outside the triangle, BD is an infinitesimal of the semicircle, and

$$\begin{aligned}\mathrm{BC} \;&=\; \mathrm{BD} - \mathrm{CD} \;=\; \mathrm{BD} - (\pi/2)\,\mathrm{FD} \qquad\qquad \text{(the last by similarity)}\\ &=\; \mathrm{BD} - (\pi/2)\,(\mathrm{EF} - \mathrm{ED})\\ &=\; \mathrm{BD} - (\pi/2)\,\big([2/\pi]a\theta - [a - a\cos\theta]\big).\end{aligned}$$

Now roll the circle a total of π – θ radians, to the (green) position at right. The dot is at P, the point of tangency at T, and angle TOP is π – θ.



The diameter at T meets AH at S, at height

$$\begin{aligned}\mathrm{TS} \;&=\; (2/\pi)\,\mathrm{AT} \qquad\qquad &\text{(by similarity)}\\ &=\; (2/\pi)\,(\text{arc PT}) \qquad &\text{(distance = arc)}\\ &=\; (2/\pi)\,a(\pi - \theta).\end{aligned}$$

It meets the horizontal from P at R, at height

$$\mathrm{TR} \;=\; a + a\cos\theta \;=\; a - a\cos(\pi - \theta).$$

By the "opposite" remark above, R is higher than S.

Extend PR to meet AH at Q. Then PQ is an infinitesimal of half the outside area, PR is an infinitesimal of the semicircle, and

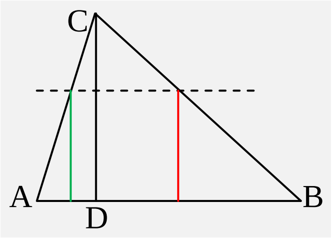
$$\begin{aligned}\mathrm{PQ} \;&=\; \mathrm{PR} + \mathrm{RQ} \;=\; \mathrm{PR} + (\pi/2)\,\mathrm{RS}\\ &=\; \mathrm{PR} + (\pi/2)\,(\mathrm{TR} - \mathrm{TS})\\ &=\; \mathrm{PR} + (\pi/2)\,\big([a - a\cos(\pi - \theta)] \;-\; [2/\pi]\,a[\pi - \theta]\big).\end{aligned}$$

The sum of the infinitesimals of the outside half-area is then

$BC + PQ = BD + PR.$ (The other terms cancel; do Exercise 3).

The equation says that the sum of the outside indivisibles equals the sum of the semicircle indivisibles. Torricelli concluded that the part of the arch above AH has the area of the semicircle. It follows that in the whole arch, the area outside the inscribed triangle equals the area of the circle. Hence the area of the arch is three circles, or 1.5 times the area of the inscribed triangle.

> [It is important that the indivisibles be "corresponding." In the semicircle, BD and PR correspond: They are symmetrically located, the same distance
>
> $OD = OR = a \cos \theta$
>
> below and above the center. **Struik** (page 102) notes that Torricelli warned Cavalieri about the following paradox: In the triangle ABC at right, for each infinitesimal (green) to the left of the altitude CD, there is an equal one (red) to the right; conclude that the altitude separates ABC into two triangles of equal area.]

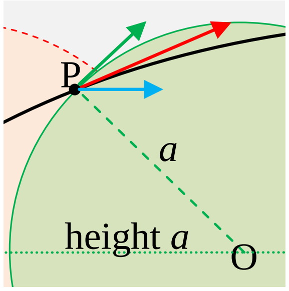


Bad year, 1647: Cavalieri died aged 49, Torricelli 39.

c) Roberval

Gilles Personne (1602-1675), called "de Roberval" after his birthplace, actually anticipated results of Torricelli, and some of Cavalieri's. For example, Roberval described the tangent to the cycloid in terms of motion.

In the cycloid figure, focus on the vicinity (magnified at right) of the tracing point P. The point is on a circle moving rightward. That motion imparts to the point a horizontal component (blue) of velocity equal to the rightward speed of the circle. At the same time, the circle is rotating. The rotation imparts a component (green) along the tangent *to the circle*.



> The speed of P along the tangent is the rate at which the arc from the bottom of the circle clockwise around to P is increasing. But the arc's length equals the horizontal distance rolled by the circle. Therefore the rate at which the arc increases is the rate at which the horizontal distance is increasing. In other words, the speed of P along the circle's tangent equals the circle's speed to the right.

Since the tangential and horizontal components are equally long, the velocity of P bisects the angle between them. That gives the direction of the (red) tangent to the cycloid: halfway between the clockwise tangent and the rightward horizontal.

You must read Boyer for the remarkable reason why Roberval did not publish his results, thereby leaving the credit for their discovery to the Italians.

Exercises VII.A.2

1. What will be the weight of a column of mercury 30 inches tall with constant horizontal cross-sections of area 1 square inch?
2. a) Adapt Torricelli's argument for the parabola to describe the tangent to an ellipse at one of its points, in terms of the focal radii to that point.
   b) What reflection property does that imply for the ellipse?
3. Verify that

$$-\Big([2/\pi]a\theta - [a - a\cos\theta]\Big) + \Big([a - a\cos(\pi - \theta)] - [2/\pi]\, a[\pi - \theta]\Big) = 0.$$

4. In our second figure above of the cycloid arch, put the x-axis of a coordinate system along the baseline AM, with the origin at A.
   a) When the circle reaches the red position, the dot at B is ($a \sin \theta$) to the left of the center and ($a \cos \theta$) lower. Show that the dot's coordinates (x, y) are given by (the "parametric equations")
   $$x = a\theta - a \sin \theta, \qquad y = a - a \cos \theta.$$
   b) Use geometry (as opposed to calculus) to express the slope of the tangent at (x, y) in terms of θ.
   c) (Calculus) Use parametric differentiation to confirm the answer in (b).
   d) Use either (b) or (c) to show that the arch does have a cusp at each end.

3. Descartes

René Descartes [day-CART] (1596-1650) was a lawyer and philosopher. Philosophy may owe more to him than math does. His *Discours sur la Méthode* … (*Discourse on the Method* [*to Reason Well and Seek Truth in the Sciences*]) tried to set forth a program by which scientific discovery could be rendered algorithmic. For our purposes, his most important contribution established a bridge between the languages of geometry and algebra.

a) the Cartesian plane

The suggestion of the Cartesian plane appeared in *La Géométrie*, now a famous book on its own, but which originally was an appendix to the *Discours*. Descartes did not develop the coordinate plane to the extent we know it now. In fact, he thought more in terms of lengths than of coordinates. Oresme's "graph" for constant acceleration (section V.B.3c) and Galileo's description of trajectories (section VI.D.3a(iii)) are closer to the way we work with coordinates and graphs. Still, his application of existing algebra to the ancient questions of geometry was a major factor in the rise of coordinate geometry.

b) Descartes's rule

One thing that came out of his work with the plane was a purely algebraic principle. We can use it to give elementary evidence to some statements for whose earlier justifications we had to invoke calculus.

[Half of] **Descartes's Rule of Signs.** A polynomial has either as many positive roots as it has changes of sign, or else an even number fewer.

Put Fibonacci's cubic equation into the **standard form**

$$x^3 + 2x^2 + 10x - 20 \;=\; 0.$$

The rule requires the powers to be in order, either decreasing or increasing. The coefficient signs are

+, + (no change), + (no change), – (change #1).

One change, the rule says, implies that the equation has one positive solution, or an even number fewer. That forces one positive solution.

Look next at

$$0 \;=\; 8 - (x - 1)^3 \;=\; 9 - 3x + 3x^2 - x^3.$$

The signs are

+, – (change #1), + (change #2), – (change #3).

The rule says that the equation has three or one positive solutions. Observe that the middle form makes clear what the rule cannot see, that $x = 3$ is the only solution. (See Exercise 1.)

For a last example, take

$$0 = (x + 3)(x - 1)^2 = x^3 + x^2 - 5x + 3.$$

From the two changes of sign, the rule predicts two or no positive solutions. From the factored form, we see that $x = 1$ is the only positive solution, but it is double. The rule counts multiplicity.

The other half of the rule addresses negative roots. Instead of writing it separately, we will replace x by $-x$ and apply the positive-roots half.

For the polynomial $p(x) = 9 - 3x + 3x^2 - x^3$, we have

$$p(-x) = 9 - 3(-x) + 3(-x)^2 - (-x)^3 = 9 + 3x + 3x^2 + x^3.$$

[Hereafter, we must make much use of function notation. Review at dummies.com, which includes how to *read* the notations $p(x)$, $p(-x)$.] In $p(-x)$, there are no changes of sign. By the positive half of the rule, $p(-x)$ has no positive roots. Therefore $p(x)$ has no negative roots (as Exercise 1 attests).

For $q(x) = x^3 + x^2 - 5x + 3$, we have

$$q(-x) = -x^3 + x^2 + 5x + 3.$$

That is one change of sign, one positive root for $q(-x)$, one negative root for $q(x)$.

Now go to our standard cubic equation,

$$0 = r(x) = x^3 + bx + c.$$

Assume first that b is positive. Then either c is positive, $r(x)$ has no change of sign, and $r(-x)$ has one; or c is negative, and the situation is left-right reversed. Either way, there is necessarily one root to one side of $x = 0$ and none to the other (Exercise 4a). Suppose instead b is negative. If c is positive, then $r(x)$ has two changes, $r(-x)$ has one. (Check that it is again opposite for negative c, so that we need not consider it separately.) In that case, a single negative root is inevitable. There are either two or no positive roots, and either number is possible (Exercise 4b).

Exercises VII.A.3

1. Factor $9 - 3x + 3x^2 - x^3$ to verify that $x = 3$ is the lone real root.
2. Judging by Descartes's Rule, how many negative roots does Fibonacci's cubic have?
3. For each equation, how many positive solutions does the rule predict, and how many negative?
   a) $x^3 - 27x - 54 = 0$. (See the analysis in section VI.B.3a.)
   b) $x^3 - 27x - 90 = 0$. (Section VI.B.3b.)
   a) $x^3 - 27x + 46 = 0$. (Section VI.B.3c.)
4. a) Sketch the two possible graphs, corresponding to positive or negative c, for
   $y = x^3 + 27x + c$.
   b) Sketch the three possible graphs of
   $y = x^3 - 27x + c$ with $c > 0$,
   one of them having no positive x-intercept, one having just one intercept, one having two. Why does the "just one" intercept not violate Descartes's Rule?

4. Fermat and the Synthesis of Algebra and Geometry

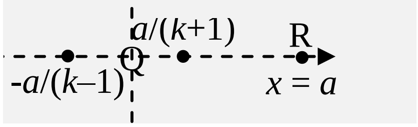
Pierre de Fermat [fair-MAH] (1601-1665) was another lawyer. He made epochal discoveries in mathematics, but treated the subject as a hobby. Almost none of his discoveries appeared in the form of books. Rather, he wrote of them to friends and mathematicians, who brought them to general attention. It was Fermat who began to cast coordinate geometry into our form (albeit largely in Quadrant I, with negative coordinates not allowed) and who created what is now analytic geometry.

a) geometry via algebra

We already studied geometric questions by applying coordinate thinking (ahead of its time). The best example is from Copernicus, the circle in the circle (section VI.D.1a). To see implications of Fermat's work on the connection suggested by Descartes, recall the ancient locus of Apollonius (from Exercise III.A.7:1).

> The locus is the set of points in a plane whose distance from one fixed point is a constant multiple of their distance from a second fixed point. Put the origin at the fixed point Q and the x-axis through fixed point R, giving it coordinates $(a, 0)$. For the generic point $P(x, y)$ to have $PR = k\ PQ$, its coordinates must satisfy
>
> $\sqrt{([x-a]^2 + [y-0]^2)} \;=\; k\,\sqrt{([x-0]^2 + [y-0]^2)}$ (distance formula).
>
> We can simplify that to
>
> $-2ax + a^2 \;=\; (k^2-1)\,x^2 + (k^2-1)\,y^2.$
>
> If $k = 1$, then the last reduces to
>
> $x = a/2.$
>
> We recognize that as the equation of the perpendicular bisector of QR.
>
> If instead k > 1—the treatment is analogous if $k < 1$—then we recast the equation as
>
> $$\begin{aligned}(x + a/[k^2-1])^2 + y^2 \quad &= \quad a^2/[k^2-1] + a^2/[k^2-1]^2\\ &= \quad a^2k^2/[k^2-1]^2.\end{aligned}$$
>
> (Verify: Exercise 1.) Since that rightmost quantity is necessarily positive, we recognize that the equation describes a circle.

We could determine the center and radius of the circle from the equation. Let us instead check geometrically, based on knowing that the locus is a circle.



> It is easy to name two points that fit the bill. In the figure at right, break up the segment QR into $k + 1$ equal parts. The first partition point is $a/(k+1)$ to the right of Q, distance
>
> $a - a/(k+1) = k\,a/(k+1)$
>
> to the left of R. Hence $(a/(k+1), 0)$ is one point on the circle. Next, look $a/(k-1)$ to the left of Q. That point is
>
> $a + a/(k-1) = k\,a/(k-1)$
>
> leftward from R. Hence $(-a/(k-1), 0)$ is also on the circle. Because the locus must be symmetric about QR, we conclude that those two points are the ends of a diameter. Therefore the center is at the midpoint, where
>
> $x \;=\; 1/2\,(a/[k+1] + -a/[k-1]) \;=\; -a/(k^2-1),$ $y = 0,$
>
> and the radius is
>
> $1/2\,(a/[k+1] - -a/[k-1]) \;=\; ak/(k^2-1).$ (Reconcile with the equation.)

Look at what we say in each language, and how we implicitly slide from one to the other. We used the distance formula, which attaches numbers to segments by means of the Pythagorean theorem. (The formula therefore depends on the axes' being perpendicular, which Descartes had not insisted on.) We used the midpoint formula, which matches the geometric notion of midpoint with the numerical notion of average. We worked the equation of the circle into a standard form related to the distance formula. We ended up connecting an equation (something algebraic) and a locus (geometric). Fermat enlarged on that connection. (Descartes had not. However, by viewing an equation in two variables as giving a relation between two *lengths*, Descartes enabled the development of the connection.)

Exercises VII.A.4a

1. Assuming $k \neq 1$, transform

$$\sqrt{([x - a]^2 + [y - 0]^2)} \;=\; k\sqrt{([x - 0]^2 + [y - 0]^2)}$$

into

$$(x + a/[k^2 - 1])^2 + y^2 \;=\; a^2k^2/[k^2 - 1]^2.$$

2. Exercise VI.A.2:2 asked for a triangle with one side 5, the altitude to that side 3, and the remaining sides in the ratio √2:1. Construct the triangle in the coordinate plane:
a) Put vertices at (0, 0) and (5, 0). Based on the text's discussion, sketch accurately the circle whose points are √2 as far from (5, 0) as they are from the origin.
b) Where on the circle is the triangle's third vertex? The answer shows that there are two noncongruent triangles that fit the description.
c) What are the (two possible pairs of) coordinates of the third vertex?

b) loci and equations

Fermat made the association between two-variable equations and loci, in a systematic way. He matched general first-degree equations

$$ax + by = c \qquad \text{(at least one of } a \text{ and } b \text{ nonzero)}$$

with lines. (Note how making "linear equation" a synonym for "first-degree equation" mixes geometric and algebraic language.) He matched general second-degree equations

$$ax^2 + bxy + cy^2 + dx + ey + f \;=\; 0$$

with conic sections. He went on to try to classify higher degrees.

To see some of the association, take the simplest quadratic graph, given by $y = x^2$. Without reference to conic sections, we can see that it has a low point at (0, 0). From there, it rises symmetrically to right and left. We habitually call it a parabola. Is it really a parabola?

> For it to be one, it must meet the parabola's characterization: Its points must be equally distant from a focus and a directrix. Since the parabola's axis would have to be the y-axis, the focus has to be $(0, f)$, some distance f above the low point. The directrix has to be the line $y = -f$, the same distance below the origin. To find f, just take another point on the locus, like (1, 1). Its distances are $1 + f$ above the line, $\sqrt{([1 - 0]^2 + [1 - f\,]^2)}$ from the proposed focus. From
>
> $$\sqrt{([1 - 0]^2 + [1 - f\,]^2)} \;=\; 1 + f,$$
>
> we conclude $f = 1/4$. Exercise 1 asks you to show that the graph of $y = x^2$ is precisely the locus of points equidistant from the point (0, 1/4) and the line $y = -1/4$. The graph really is a parabola.

In similar fashion, we classify the graph of $y = 1/x$ as a hyperbola. (Writing $y = x^{-1}$ allows us to think of it as a power graph; writing $xy = 1$ puts it into the second-degree family.) Does it meet the criterion?

> For a hyperbola, the transverse axis bisects the angle between the asymptotes. Hence the 45°-line would have to be the axis, and the points (1, 1) and (-1, -1) the two vertices. Their separation is $2\sqrt{2}$, and that would be the constant difference between any given point's focal radii. The foci would have to be on the axis, say at $(f, f\,)$ and $(-f, -f\,)$.
>
> We can find f by algebra, beginning with
>
> $$[\text{distance from } (2, 1/2) \text{ to } (f, f\,)] \;=\; [\text{distance from } (2, 1/2) \text{ to } (-f, -f\,)] - 2\sqrt{2}.$$
>
> Too much work: Let us instead recall that if a hyperbola is **rectangular** (asymptotes perpendicular), then its foci are √2 as far from the center as the vertices are. That puts the foci at (√2, √2) and (-√2, -√2). Exercise 2 asks you to verify that our graph is indeed a hyperbola.

Those uncomplicated examples belie the difficulty in linking second-degree equations with conics. (The difficulty is there without even considering their degenerate forms: $xy = k$ gives a hyperbola whether k is positive or negative, but not if $k = 0$.) Remember that turning

$$ax^2+ bxy + cy^2 + dx + ey + f = 0$$

into a standard form requires rotation of the axes, which Fermat demonstrated. (Compare Exercise 3.)

Exercises VII.A.4b

1. a) Show that if a point (x, y) is equally distant from the point $(0, 1/4)$ and the line $y = -1/4$, then it satisfies $y = x^2$.
   b) Show conversely that if a point (x, y) satisfies $y = x^2$, then it is equidistant from the point $(0, 1/4)$ and the line $y = -1/4$.
2. a) Show that if the distance from (x, y) to $(\sqrt{2}, \sqrt{2})$, minus the distance from (x, y) to $(-\sqrt{2}, -\sqrt{2})$, is $\pm 2\sqrt{2}$, then (x, y) satisfies $y = 1/x$. (This is a variation on a standard demonstration in textbooks covering analytic geometry.)
   b) Show conversely that if (x, y) satisfies $y = 1/x$, then its distances to $(\sqrt{2}, \sqrt{2})$ and $(-\sqrt{2}, -\sqrt{2})$ differ by $\pm 2\sqrt{2}$.
3. Suppose you set up the vw-coordinate system at the same origin as the xy-system, with the v-axis going up to the right, the w-axis up to the left, at 45° angles to the xy-axes. Then (we accept that) the coordinates (v, w) of a point relate to the (x, y) coordinates by
   $$v = x\sqrt{2}/2 + y\sqrt{2}/2, \qquad w = -x\sqrt{2}/2 + y\sqrt{2}/2.$$
   a) Solve those equations for x and y in terms of v and w.
   b) Substitute the result of (a) to turn $xy = 1$ into (the now standard form of hyperbola)
   $$v^2/2 - w^2/2 = 1.$$

---

c) algebra via geometry

We applied coordinate geometry to algebra as early as section IV.B.3, on Zhu Shijie's ("Horner's") method. Now we draw conclusions about the solutions of cubic equations from the geometry of graphs.

In Section VI.B, studying cubics algebraically, we reduced the most general cubic to the form

$$x^3 + bx + c = 0.$$

Assume that c is negative; if it is positive, the argument needed is the mirror-image of that below. We may then write $c = -a^2/2$ and rewrite the equation as

$$x^2 + b = a^2/2x.$$

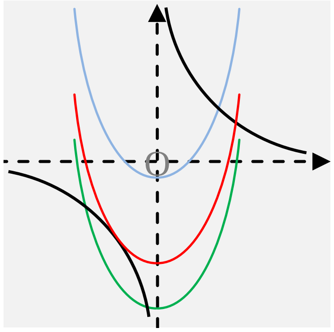
Therefore the solutions to the cubic are given by the intersections of the two graphs

$$y = x^2 + b \qquad \text{and} \qquad y = a^2/2x.$$

(i) the possibilities

> Adapting the discussion from (b), we see that the graph of $y = x^2 + b$ is a parabola with low point at $(0, b)$, focus $(0, b + 1/4)$, directrix $y = b - 1/4$. The graph of $y = a^2/2x$ is a hyperbola with vertices at $\pm(a/\sqrt{2}, a/\sqrt{2})$, which imply foci at $\pm(a, a)$. The focal radii from any of its points differ by the distance between those vertices, $2a$.

The hyperbola has origin-symmetric halves, black in the figure below, in Quadrants I and III. The Quadrant I half starts indefinitely high indefinitely close to the y-axis, then drops indefinitely to the right to approach indefinitely close to the x-axis. That means it starts above and left of the parabola, ends up below and right, no matter what b and a are. It is intuitively clear—without reference to theorems from calculus —that the graphs must meet in the quadrant. Also, they diverge to left and right of the meeting, so the intersection is unique. We conclude that the original cubic has exactly one positive solution.

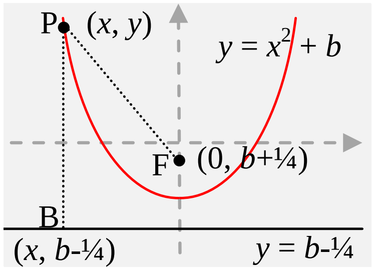


If b is positive or not too negative, then the left half of the parabola (blue curve in the figure) is entirely above the Quadrant III part of the hyperbola (black). In that case, the cubic has no negative solutions. If b is sufficiently below zero, then the left half of the parabola is low enough (green) to cross the hyperbola twice in the quadrant. That situation implies that the cubic has two negative solutions to go with the one positive.

Somewhere between no intersection and two intersections, we have the possibility that the two graphs have a single common point in Quadrant III. That means the parabola (red in the figure) is tangent to the hyperbola. We will look for common tangents, to characterize the situation when there is a lone (double) negative solution. (Later we will discuss Fermat's algebraic characterization of the tangents, but here we call upon the geometry of Apollonius and Torricelli.)

(ii) tangent to the parabola

Remember Torricelli's statement (section VII.A.2a) that the tangent at point P bisects the angle between the line (PF in the figure at right) to the focus and the perpendicular (PB) to the directrix. We fix the tangent in terms of slopes.



By the definition of the parabola, PF and PB are equal distances. Therefore the base BF of the isosceles triangle PBF is perpendicular to the bisector of the angle at P. The slope of BF is

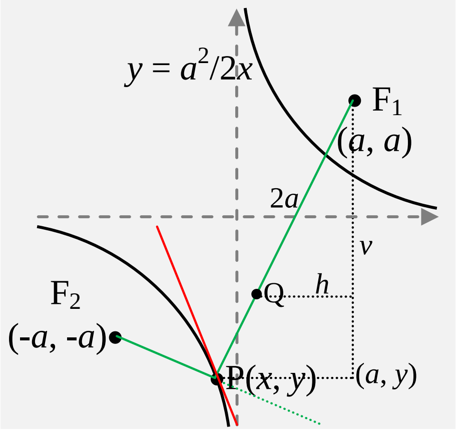
$$(b + 1/4 - [b - 1/4])/(0 - x) \;=\; -1/2x,$$

assuming P is not at the vertex $(0, b)$. We conclude that the slope of the tangent at P is $2x$.

Notice that the slope happens to be $2x$ even when P *is* at the vertex and $x = 0$.

(iii) tangent to the hyperbola

Our two sources say that the tangent to a hyperbola bisects the angle between the focal radii.



In the figure at right, we see the focal radii (solid green) from $P(x, y)$ to $F_1(a, a)$ and $F_2(-a, -a)$, as well as the (red) tangent at P to the hyperbola. Put Q along PF_1 to make $PQ = PF_2$. Triangle PQF_2 is isosceles. Hence the tangent, bisector of the angle at P, is perpendicular to QF_2.

We need to spot Q. Begin with the fact that

$$2a \;=\; PF_1 - PF_2 \;=\; PF_1 - PQ \;=\; QF_1.$$

Assume that Q is h to the left and v down from F_1. From the right triangles with dotted black legs, we see that

$$h/2a = (a - x)/PF_1 \qquad \text{and} \qquad v/2a = (a - y)/PF_1.$$

Therefore the coordinates of Q are

$$x_Q \;=\; a - h \;=\; a - 2a(a - x)/PF_1,$$
$$y_Q \;=\; a - v \;=\; a - 2a(a - y)/PF_1.$$

We then conclude that the slope of QF_2 is

$$[a - 2a(a - y)/PF_1 + a]/[a - 2a(a - x)/PF_1 + a]$$
$$= \qquad (2aPF_1 - 2a^2 + 2ay)/(2aPF_1 - 2a^2 + 2ax).$$

Consequently the tangent at P has slope

$$-(2aPF_1 - 2a^2 + 2ax)/(2aPF_1 - 2a^2 + 2ay).$$

Write the relation $\mathrm{PF}_2 = \mathrm{PF}_1 - 2a$ as

$$\sqrt{([x + a]^2 + [y + a]^2)} \quad = \quad \sqrt{([x - a]^2 + [y - a]^2)} - 2a.$$

Square to the form

$$([x + a]^2 + [y + a]^2) \quad = \quad ([x - a]^2 + [y - a]^2) - 4a\,\mathrm{PF}_1 + 4a^2,$$

and simplify to find

$$\mathrm{PF}_1 = -x - y + a.$$

Then we can substitute to rewrite the slope of the tangent at P(x, y) as

$$-(2a[-x - y + a] - 2a^2 + 2ax)/(2a[-x - y + a] - 2a^2 + 2ay)$$
$$= (2ay)/(-2ax) \;=\; -a^2/2x^2.$$

[Yes, you do have to check that simplification. Remember that $y = a^2/2x$, with x never zero.]

(iv) the common tangents

For the parabola and hyperbola to be tangent at (x, y), we need

$$2x \;=\; -a^2/2x^2, \qquad \text{forcing}$$
$$x \;=\; \sqrt[3]{(-a^2/4)} \;=\; -\sqrt[3]{(a^2/4)}.$$

In terms of the original $c = -a^2/2$, that means the cubic's solution is at $x = -\sqrt[3]{(-c/2)}$. That finding agrees with the section VI.B.3a placement of the double root. Because the parabola and hyperbola have the corresponding point in common, we know

$$y \quad = \quad x^2 + b \quad = \quad (-c/2)^{2/3} + b \quad = \quad c^{2/3}2^{-2/3} + b$$

must be the same as

$$y \quad = \quad -c/x \quad = \quad c(-c/2)^{-1/3} \quad = \quad -c^{2/3}2^{1/3}.$$

Therefore the original cubic must have had

$$b \;=\; -c^{2/3}2^{-2/3} - c^{2/3}2^{1/3} \;=\; -c^{2/3}2^{-2/3}[1 + 2], \qquad \text{or}$$
$$b^3 \;=\; -c^2 2^{-2}[3]^3 \;=\; -27c^2/4.$$

It must have had the discriminant $c^2 + 4b^3/27$ equal to zero.

Except for our use of coordinates (and therefore slopes), this argument is the kind Omar Khayam brought to bear in his analysis of cubics. See the end of Section V.A.3, especially Exercise 5 there.

Exercises VII.A.4b

1. Recall Apollonius's result (section III.A.6a(iii)) that on a parabola with a given chord, the place where the tangent is parallel to the chord is halfway (horizontally in our situation) between the ends of the chord. Pick a point of the parabola $y = x^2 + b$ to the right of (x, y) and a second point equally far left. Find the slope of the chord joining the points, to verify that the slope of the tangent at (x, y) is $2x$.

d) tangents to power graphs

Fermat's algebraic (coordinate) geometry gave him the tangents to the whole class of power graphs. His method is precisely the way we introduce tangents for polynomial graphs in beginning calculus.

Let us work the example $y = x^4$; the extension to general positive integer powers will be obvious.

To describe the tangent at the point (a, a^4), Fermat took the nearby point ($[a + h]$, $[a + h]^4$), with some small positive number h. He examined the slope of (what we call) the **secant** joining them. That would be

$$([a + h]^4 - a^4)/([a + h] - a) \;=\; (4a^3h + 6a^2h^2 + 4ah^3 + h^4)/h$$
$$=\; 4a^3 + 6a^2h + 4ah^2 + h^3.$$

Setting $h = 0$ (Fermat could not have thought in terms of "limits") makes the fraction meaningless, but not the last expression. It becomes $4a^3$, which Fermat was happy to call the slope of the tangent.

The method works for negative exponents also; see Exercise 1. For all integers, the tangent to the graph of $y = x^n$ at (a, a^n) has slope na^{n-1}.

This method arose, not in the investigation of tangents per se, but in locating the turning points of polynomial graphs. In the book *Method* [*for*] *Finding Maxima and Minima* (published, of course, after Fermat died), he described the equivalent of what we sometimes call Fermat's theorem: The maxima and minima of polynomials are to be found among the places where the slope of the tangent is zero. (See Exercises 2 and 3.)

Exercises VII.A.4d

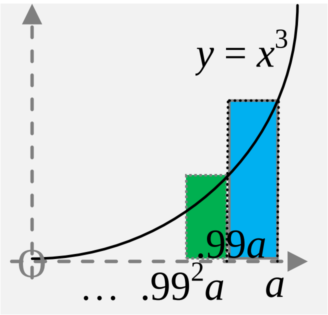
1. Apply Fermat's method to find the slope of the tangent to $y = x^{-3}$ at the point (a, a^{-3}).
2. a) Apply the method to $y = x(10 - x)$.
   b) Use the slope to find the maximum possible value of $x(10 - x)$. (Compare Exercise VI.B.4b:1.)
3. a) Apply the method to $y = x^3 - 27x - 54$.
   b) Use the slope to find the graph's high point to the left of the y-axis, and the low point to the right.

e) areas under power graphs

(i) positive powers

Fermat proceeded to the areas under power graphs. His method there had elements of our approach in elementary calculus, but used an ingenious twist.

Consider the region under the graph of $y = x^3$ from $x = 0$ to $x = a$, pictured at right. Partition it into vertical strips. From right to left, make each strip cover the same fraction *of the width not yet taken*.



For example, use the fraction 0.01. The first strip spans the rightmost 1% of the region, the part lying between $x = .99a$ and $x = a$. At the right edge of the strip, the graph has height a^3. Therefore its circumscribed rectangle (filled in blue in the figure) has area

$$(a - .99a)\, a^3 = .01\, a^4.$$

The second strip spans 1% of what remains from $x = 0$ to $x = .99a$. That is, it covers $x = .99(.99a) = .99^2a$ to $x = .99a$. At its right edge, the graph has height $(.99a)^3$. Accordingly, its circumscribed rectangle (green) has area

$$(.99a - .99^2a)\,(.99a)^3 = (.99^4)\ .01\, a^4.$$

For the third strip, the span is $x = .99^3a$ to $x = .99^2a$. That implies a rectangle with area

$$(.99^2a - .99^3a)\,(.99^2a)^3 = (.99^8)\ .01\, a^4. \qquad \text{(Check the algebra.)}$$

The pattern is clear.

The areas of this infinity of rectangles add up to

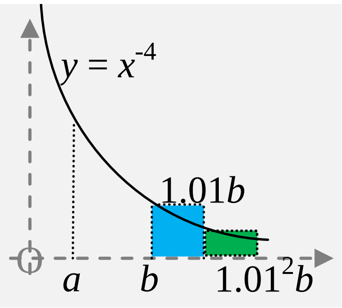
$$\begin{aligned} &.01\, a^4 + (.99^4)\ .01\, a^4 + (.99^8)\ .01\, a^4 + \ldots \\ &= \quad .01\, a^4\, [1 + .99^4 + .99^8 + \ldots] \\ &= \quad (1 - .99)\, a^4\, 1/(1 - .99^4) \\ &= \quad a^4/(1 + .99 + .99^2 + .99^3). \qquad \text{(Explain them all!)} \end{aligned}$$

Now replace the fraction 0.01 by 0, pushing .99 to 1.00. That shrinks the rectangles to infinitesimals that exactly cover the region under the graph. We judge the region's area to be $a^4/4$.

Our rectangles obviously overestimate the area: They use the maximal height, on the right. Fermat does not seem to have shared Archimedes's wish to squeeze the area between the "upper estimate" and the lower. Do Exercise 1 to see that using the minimal heights, on the left, leads to the same area.

(ii) negative powers

With negative exponents, Fermat argued even more cleverly. He turned the question inside out: To study the region under $y = x^{-4}$ from $x = a$ to $x = b$, examine instead the infinitely long region to its right.



Partition the part *rightward* of $x = b$ geometrically ("fractionally"). Using 1% again, we put the breaks at $x = 1.01b$, $x = 1.01^2b$, …. Out there, the first strip (blue in this figure) extends from $x = b$ to $x = 1.01b$. Consequently the circumscribed rectangle has width $(1.01b - b)$, height b^{-4} on the left, area

$$(1.01b - b)\, b^{-4} \quad = \quad .01\, b^{-3}.$$

The next rectangle (green) has width $(1.01^2b - 1.01b)$, height $(1.01b)^{-4}$ on the left, area

$$(1.01^2b - 1.01b)\,(1.01b)^{-4} \quad = \quad (1.01^{-3})\, .01\, b^{-3}.$$

For the one after that, it is base $(1.01^3b - 1.01^2b)$, height $(1.01^2b)^{-4}$, area

$$(1.01^3b - 1.01^2b)\,(1.01^2b)^{-4} \quad = \quad (1.01^{-6})\, .01\, b^{-3}.$$

(Check the last two!) We conclude that the rectangle areas add up to

$$\begin{aligned} &.01\, b^{-3} + (1.01^{-3})\, .01\, b^{-3} + (1.01^{-6})\, .01\, b^{-3} + \ldots \\ &\quad = \quad .01\, b^{-3}\, [1 + 1.01^{-3} + 1.01^{-6} + \ldots] \\ &\quad = \quad b^{-3}\, (1.01 - 1)\, 1/(1 - 1.01^{-3}) \\ &\quad = \quad b^{-3}\, 1.01^3/(1.01^2 + 1.01 + 1). \qquad \text{(Explain.)} \end{aligned}$$

Replace 1.01 by 1.00 to conclude that the area beyond $x = b$ is $b^{-3}/3$.

The area beyond $x = a$ must be $a^{-3}/3$. Therefore the area between $x = a$ and $x = b$ is $a^{-3}/3 - b^{-3}/3$. [In elementary calculus, we would write it as $b^{-3}/\text{-}3 - a^{-3}/\text{-}3$.]

(iii) the special case

The argument in (i) for $y = x^3$ adapts easily to rational exponents (Exercise 2), including negative fractions exceeding -1. The argument for $y = x^{-4}$ (ii) adapts to negative fractions below -1 (Exercise 3). However, each argument lands on its face if the exponent is -1. For the graph of $y = x^{-1}$, the crucial sums (the sums in red) are infinite, either

$$[1 + .99^0 + .99^0 + \ldots] \quad \text{or} \quad [1 + 1.01^0 + 1.01^0 + \ldots].$$

Still, the specific case $y = x^{-1}$ has a peculiar interest. It ends up relating to logarithms.

Apply Fermat's scheme to the graph of $y = 1/x$ from $x = a$ to $x = b$. The strip with the rightmost 1% of the region starts at $x = a + .99(b - a)$ and ends at $x = b$. Of the remaining expanse from $x = a$ to $x = a + .99(b - a)$, the rightmost 1% starts at $x = a + .99^2(b - a)$ and ends at $x = a + .99(b - a)$. The third strip covers $x = a + .99^3(b - a)$ to $x = a + .99^2(b - a)$, and we see the pattern.

With those endpoints, the strips have widths

$$\begin{aligned} b - [a + .99(b - a)] &\quad = \quad .01\,(b - a), \\ [a + .99(b - a)] - [a + .99^2(b - a)] &\quad = \quad .99\,(.01)\,(b - a), \\ [a + .99^2(b - a)] - [a + .99^3(b - a)] &\quad = \quad .99^2\,(.01)\,(b - a), \ldots. \end{aligned}$$

Using the minimal heights, at the endpoints on the right, we establish *inscribed* rectangles of heights

$$1/b, \qquad 1/[a + .99(b - a)], \qquad 1/[a + .99^2(b - a)], \ldots.$$

Therefore the areas of the rectangles add up to

$$.01(b - a)/b + (.99).01(b - a)/[a + .99(b - a)] + (.99)^2.01(b - a)/[a + .99^2(b - a)] + \ldots.$$

Those terms are not in any familiar progression. We have no easy way to sum them. However, the sum has one reasonable feature: It has *a* and *b* in first degree throughout, with no powers or roots.

> We can divide all the numerators and denominators by *a*. Thereby the sum becomes
>
> $$.01(b/a - 1)/(b/a) + (.99).01(b/a - 1)/[1 + .99(b/a - 1)] + (.99)^2.01(b/a - 1)/(1 + .99^2(b/a - 1) + \ldots.$$
>
> Observe that the .01’s and .99’s stayed put. This new sum has the same structure as the previous, except that *b*/*a* replaced *b* and 1 replaced *a*. In other words, the 1% approximation for the area from $x = a$ to $x = b$ is the same as the 1% approximation would be for the area from $x = 1$ to $x = b/a$. The conclusion is inescapable: The area from $x = a$ to $x = b$ equals the area from $x = 1$ to $x = b/a$.

Use the function notation $L(t)$ to denote the area under $y = 1/x$ from $x = 1$ to $x = t$. The area from $x = 1$ to $x = b/a$ is $L(b/a)$. We just argued that this matches the area from $x = a$ to $x = b$. Draw a figure to convince yourself that the latter area is $L(b) - L(a)$. The function *L* has the property that

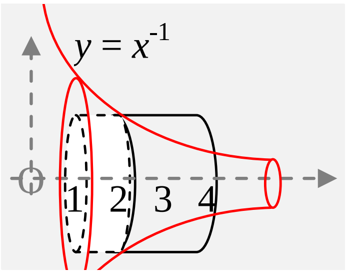
$$L(b/a) = L(b) - L(a).$$

Among our functions, the only kind that turns division into subtraction that way is the logarithm. Let us accept that $L(t)$ is the logarithm of *t* to some base, which we can only estimate; see Exercise 4.

[Bear in mind that Fermat would not have thought this way. Logarithms were invented around 1600 as *devices for multiplication*. Except in some work by Torricelli (see Boyer), the idea of a log *function* developed after Fermat. It had been like that with trigonometry. Trigonometry was a measurement tool for thousands of years, then under Viète a device for calculation (Boyer). It was long after Fermat that the idea of trigonometric functions came under study.]

Exercises VII.A.4e

1. Referring to the figure in (i) of the graph of $y = x^3$: Show that if we use the minimal height (at the left) of each strip to approximate the total area, then we are still led to the area $a^4/4$. (Shortcut: By exactly how much does the upper estimate exceed the lower?)
2. Try the area argument for $y = x^{5/2}$. You will easily find that the rectangles sum up to
   $$a^{7/2}\ (1 - .99)\ 1/(1 - .99^{7/2}).$$
   To factor numerator and denominator, set $c = .99^{1/2}$. Simplify the fraction, replace *c* by 1, and lead to area = $a^{7/2}/(7/2)$.
3. Use the “area rightward” argument to evaluate the area under the graph of $y = x^{-5/2}$ between $x = a$ and $x = b$.
4. When we let $L(t)$ denote the area under $y = 1/x$, we concluded that $L(t) = \log t$, to an unnamed base. Leonhard Euler gave that base the designation “*e*.” Then $L(e) = \log_e e$ has to be 1. Sketch the graph of $y = 1/x$ from $x = 1$ to $x = 3$, approximate areas under it in some simple way, and use the approximation to estimate *e*.
5. Suppose we revolve the region under $y = 1/x$, rightward from $x = 1$, about the *x*-axis. We produce an infinitely long solid of revolution. The figure below right shows some of it, outlined in red. It is nicknamed “Torricelli’s trumpet,” because he found the results below.



a) In the figure, the dotted black outline is the inscribed cylinder between $x = 1$ and $x = 2$. The surface area of the trumpet surrounding it is clearly greater than the curved area of the cylinder. (Think of the trumpet as consisting of infinitesimal round bands of bigger diameter than the cylinder’s.) Show that the surface area of the whole trumpet is infinite, by summing the areas of the inscribed cylinders from 1 to 2, 2 to 3, …. (Hint: You need Oresme’s deduction, section V.B.3b, that the harmonic

series adds up to infinity.)
b) The solid black outline is the circumscribed cylinder between $x = 2$ and $x = 4$. Clearly the volume of the contained part of the trumpet is smaller than the volume of the cylinder. Show that the whole trumpet has finite volume, by summing the volumes of the circumscribed cylinders from 2 to 4, 4 to 8, …. [This was a strange finding: a finite volume, despite the solid's infinite extent. I once heard this description: You can fill the trumpet with a finite amount of paint, but you cannot paint its inside.]

f) number theory

Fermat's number theory is worth a detour off the road to calculus. He had a knack for discerning number patterns. Here we study four that inspired fruitful research and beautiful results in the theory.

(i) Fermat's Little Theorem

Theorem 1. If p is a prime that does not divide a, then $a^{p-1} - 1$ is divisible by p.

Fermat extrapolated the statement from many examples. We could give his elementary proof, but it would be clumsy with our current arithmetical language. Later, we will write two nice proofs, an elementary one from Euler and an elegant one in the language introduced by Carl Gauss. ["Our current arithmetical language" sometimes forces us to use remainders. We could put the theorem thus: If p is a prime that does not divide a, then a^{p-1} has remainder 1 on division by p.]

Look at examples.

Let $p = 7$. We have

$1^6 - 1 = 0, \qquad 2^6 - 1 = 63 = 9 \times 7, \qquad 3^6 - 1 = 728 = 104 \times 7.$

Instead of calculating 4^6, 5^6, 6^6, observe that the binomial theorem gives

$$4^6 - 1 = (7-3)^6 - 1 = 7^6 - 6(7)^5 3 + \ldots - 6(7)^1 3^5 + 3^6 - 1.$$

The red terms are all multiples of 7, and we already know that $3^6 - 1$ is another multiple. Therefore

$$4^6 - 1 = (\text{multiplier}) \times 7,$$

and analogously with 5 and 6.

The theorem is not confined to the numbers below p. We have

$$\begin{aligned} 60^6 - 1 &= (8 \times 7 + 4)^6 - 1 \\ &= (8 \times 7)^6 + 6(8 \times 7)^5 4 + \ldots + 6(8 \times 7)^1 4^5 + 4^6 - 1 \\ &= \text{multiple of 7} + \text{another multiple of 7}. \end{aligned}$$

Checking is harder with bigger primes, but we can use remainders here and there to save calculation.

Take $p = 43$, $a = 9$. For the powers of 9, write first

$$9^2 = 43 + 38.$$

Let us stretch "remainder" and write

$$9^2 = 2 \times 43 - 5.$$

Then

$$\begin{aligned} 9^4 &= (i)43 + 25, \\ 9^8 &= (j)43 + 625 = (j + 14)43 + 23, \\ 9^{16} &= (k)43 + 529 = (k + 12)43 + 13, \\ 9^{32} &= (l)43 + 169 = (l + 3)43 + 40 = (l + 4)43 - 3. \end{aligned}$$

(We are as willing to tweak the division algorithm of the Greeks [section III.B.1b] as we are to adapt the multiplication-by-doubling of the Egyptians [section II.A.2].) Finally,

$$\begin{aligned} 9^{42} = 9^{32}\, 9^8\, 9^2 &= ([l + 4]43 - 3)\,([j + 14]43 + 23)\,(2 \times 43 - 5) \\ &= (m)43 + (-3)(23)(-5) = (m + 8)43 + 1. \end{aligned}$$

Therefore $9^{42} - 1$ is divisible by 43. (Gauss's language will make such calculations much easier.)

(ii) sum and difference of squares

Proposition. If p is a prime whose remainder on division by 4 is 1, then there is a unique pair of natural numbers whose squares add up to p.

Again, Fermat generalized from a multitude of examples. For this statement, he did not have a proof. The first known proof was by Euler.

To give examples:

$13 = 3^2 + 2^2, \quad 61 = 6^2 + 5^2.$

The theorem is silent about 21 and 65, because they are composite. You can see that no two squares sum to 21, whereas $65 = 7^2 + 4^2 = 8^2 + 1^2$. It is likewise silent on primes whose remainder is 3. We will confirm later that no such number, prime or not, can be the sum of two squares. (Compare Exercise II.B.1:1a-b)

Boyer says that Fermat knew that any prime is the *difference* of just one pair of squares. We can improve on that with the following theorem.

Theorem 2. A number can be written as the difference of nonnegative squares in as many ways as it can be written as the product of two factors of the same parity. In symbols: Given a natural k, the number of pairs $m \geq n \geq 0$ with $k = m^2 - n^2$ equals the number of pairs $u \geq v \geq 1$ of the same parity with $k = uv$.

For an example, remember that the j'th and $(j - 1)$'th squares differ by the j'th odd number:

$j^2 - (j-1)^2 = 2j - 1.$

That means every odd number is the difference of consecutive squares. Thus,

$21 = 11^2 - 10^2, \qquad 47 = 24^2 - 23^2.$

We know 47 × 1 is the only factoring of 47. By Theorem 2, we infer that 24^2 and 23^2 form the only pair whose difference is 47. On the other hand, $21 = 21 \times 1 = 7 \times 3$. That is two factorings; there must exist some other pair of squares whose difference is 21. (What is that other pair?)

Similarly, if k is double an odd number, the way 2 and 62 are, then $k = m^2 - n^2$ is impossible. In this case, in any factorization of k, one factor has the 2 and the other does not. In other words, the two factors are necessarily of opposite parity; no squares differ by k.

To prove the theorem, we will match up differences of squares and same-parity factorizations.

Suppose k has a difference-of-squares expression

$k = m^2 - n^2, \qquad m \geq n \geq 0.$

The two natural numbers

$u = m + n \quad$ and $\quad v = m - n$

differ by the even number $2n$. They therefore have the same parity, and clearly $k = uv$. In that manner, each expression of k as difference of squares yields an expression of k as the product of two factors of like parity.

Different difference-of-squares expressions yield different like-parity factorizations. Suppose k has a second difference-of-squares expression

$k = M^2 - N^2, \qquad$ with say $M > m$.

Let

$U = M + N, \qquad V = M - N.$

Then $k = UV$ has to be a *new* same-parity factoring. We cannot have $U = v$, because $U \geq M > m \geq v$. We also cannot have $U = u$, because that would force $V = v$, so that

$M = (U + V)/2 \quad$ would equal $\quad (u + v)/2 = m$.

Thus, each difference yields a factoring, and different differences yield different factorings. It follows that there are at least as many factorings as there are differences. (See also Exercise 3.)

Finally, every factoring comes from a difference. If $k = \mathcal{U}\mathcal{V}$ is such a factoring, then each of

$$\mathcal{M} = (\mathcal{U} + \mathcal{V})/2 \qquad \text{and} \qquad \mathcal{N} = (\mathcal{U} - \mathcal{V})/2$$

is a nonnegative integer, because $\mathcal{U} + \mathcal{V}$ and $\mathcal{U} - \mathcal{V}$ are both even. We have

$$k = \mathcal{M}^2 - \mathcal{N}^2 \qquad \text{and} \qquad \mathcal{U} = \mathcal{M} + \mathcal{N}, \quad \mathcal{V} = \mathcal{M} - \mathcal{N}. \qquad \text{(Verify!)}$$

Those say that $\mathcal{U}\mathcal{V}$ comes from $\mathcal{M}^2 - \mathcal{N}^2$. Since every factoring comes from a difference, and each difference yields just one factoring, there must be at least as many differences as there are factorings.

Doubtless you can see the **pigeonhole principle** at work in that argument. In basic form, it says that if you try to put $n + 1$ objects into n pigeonholes, then at least one pigeonhole will get more than one object. [I don’t know if the name comes from actual birds, or from the British use of “pigeonhole” for *mailbox*. If you try to stuff $n + 1$ letters into n mailboxes, ….] It implies that if you stuff letters into mailboxes, different letters in different boxes, using up all the boxes, then the number of letters must have been the same as the number of mailboxes. That is how we matched differences and factorings.

(iii) Fermat’s primes

Fermat guessed that every number of the form $2^{2^n} + 1$ is prime. We verify the first three examples immediately: $2^{2^0} + 1 = 3$, $2^{2^1} + 1 = 5$, $2^{2^2} + 1 = 17$.

To check

$$2^{2^3} + 1 = 256 + 1 = 257,$$

we need only divide by primes $\sqrt{257} = 16+$ or less. None of 2, 3, 5, 7, 11, 13 divides 257; it is prime. The next one,

$$2^{2^4} + 1 = 2^{16} + 1,$$

happens to be prime, but takes long to check. Remember that $2^{10} \approx 10^3$. That means

$$2^{16} + 1 = 2^6\, 2^{10} + 1 \approx 64{,}000.$$

The square root of that number is $2^8+ = 256+$. That makes for a good few primes to try.

Fermat was mistaken. It turns out that

$$2^{2^5} + 1 = 2^{32} + 1 \approx 4 \times 10^9$$

is composite. But no human could have tried dividing by all the primes below its square root, just over 2^{16}. A better approach was required. [Guess who provided it.]

(iv) Fermat’s Last Theorem

Fermat guessed that the Diophantine equation

$$x^n + y^n = z^n$$

has no natural-number solutions if $n \geq 3$. The statement is called “Fermat’s Last Theorem.” Remember that we characterized the solutions for $n = 2$, namely the Pythagorean triples. Fermat took care of $n = 4$; he proved that $x^4 + y^4 = z^4$ is impossible. In fact, he proved that you cannot solve even $x^2 + y^4 = z^4$.

Theorem 3. The equation

$$x^2 + y^4 = z^4$$

has no natural solutions.

Assume $x = A$, $y = B$, $z = C$ is a solution. Let d be the greatest divisor of B and C. Then d^4 divides $C^4 - B^4 = A^2$. It follows that d^2 divides A (Exercise 4a). From

$$(A/d^2)^2 + (B/d)^4 = (C/d)^4,$$

we see that $a = A/d^2$, $b = B/d$, and $c = C/d$ make

$$a^2 + b^4 = c^4.$$

We have produced another solution, one in which b and c are relatively prime and the numbers are no bigger than before.

Because

$$a^2 + (b^2)^2 = (c^2)^2,$$

we know a, b^2, and c^2 form a primitive Pythagorean triple. Therefore (Theorem 1 in section II.B.1) there exist relatively prime u and v, of opposite parity, such that one of a and b^2 is $u^2 - v^2$, the other is $2uv$, and $c^2 = u^2 + v^2$. (Remember that a and b do not have symmetric roles.)

Suppose first that

$$a = 2uv, \qquad b^2 = u^2 - v^2. \qquad \text{Then}$$
$$b^2c^2 = (u^2 - v^2)(u^2 + v^2) = u^4 - v^4, \qquad \text{or}$$
$$(bc)^2 + v^4 = u^4.$$

We have another solution, and it has a smaller number on the right: $u^4 = (u^2)^2 < (c^2)^2$.

Suppose instead that

$$a = u^2 - v^2, \qquad b^2 = 2uv.$$

The product $2uv$ is a square, with u and v relatively prime. That can happen only one way (Exercise 4b): In its prime factorization, one of u and v has an odd number of 2's and even numbers of all its other prime factors; and the other has even numbers of primes, all those primes odd and different from the previous. In other words, there are relatively prime m and n, the latter odd, such that $u = 2m^2$ and $v = n^2$, or vice-versa. Therefore

$$c^2 = u^2 + v^2 = (2m^2)^2 + (n^2)^2.$$

We now have $2m^2$, n^2, and c forming a primitive triple. That means there exist relatively prime U and V such that

$$2m^2 = 2UV, \qquad n^2 = U^2 - V^2, \qquad c = U^2 + V^2.$$

(There is no question which one, $2m^2$ or n^2, is even.) From $m^2 = UV$, we conclude (Exercise 4c) that each of U and V is a square, say $U = s^2$, $V = t^2$. The middle equation has

$$n^2 = (s^2)^2 - (t^2)^2, \qquad \text{or}$$
$$n^2 + t^4 = s^4.$$

We have another solution whose right side is smaller than in the previous: $s^4 = U^2 < c < c^4$.

In the detailed algebra above, nothing is by itself contradictory. It merely says that if you name one solution, then it is possible to construct another with a smaller right side. But a contradiction is inherent there. The contradiction underlies the method of "infinite descent," which Fermat used often. The argument above is **recursive**; you can apply it repeatedly. Accordingly, one solution to

$$x^2 + y^4 = z^4$$

leads to a smaller one, which leads to a smaller one, *ad infinitum*. That is impossible: You cannot descend indefinitely through the natural numbers. Therefore no solution is possible.

Theorem 3 has an odd geometric consequence: The area of a right triangle with integer sides cannot be a square.

Suppose a right triangle has sides a, b, and c, and its area $ab/2$ is j^2. Write

$$a^2 = c^2 - b^2,$$

and let d be the GCD of b and c. Then d divides a (Exercise 4d), and

$$(a/d)^2 + (b/d)^2 = (c/d)^2$$

names a primitive triple. Its corresponding triangle's area $(a/d)(b/d)/2 = j^2/d^2$ is still a square.

We again have relatively prime u and v with (say)

$$a/d = u^2 - v^2, \qquad b/d = 2uv, \qquad c/d = u^2 + v^2.$$

Then

$$(j/d)^2 = (a/d)(b/d) = (u^2 - v^2)\,uv.$$

Each of u and v is relatively prime to $u^2 - v^2$. After all, any divisor of both u and $u^2 - v^2$ would divide $u^2 - (u^2 - v^2) = v^2$; and similarly with v. By Exercise 4c, each of $u^2 - v^2$, u, and v has to be a square. But then, $u = k^2$, $v = l^2$, $u^2 - v^2 = m^2$ give

$$m^2 = u^2 - v^2 = k^4 - l^4.$$

That contradicts Theorem 3. The area cannot be a square.

Fermat had the "Last Theorem" right, but it took the world of mathematics 350 years to prove it. (Boyer, writing in 1968, said "the problem remains unsolved." **Merzbach**, on page 328 from 2011, writes "the problem remained unsolved until the 1990s."[sic]) The statement's relative elementariness and durable resistance to proof would by themselves have made it one of the most famous number conjectures of all time. What made it irresistible was that Fermat *said* he proved it, presumably easily. (Remember that Fermat's "easily" included arguments like Theorem 3's above.) Fermat had the habit of writing in the margins of his books. He made the statement—in his language, "to divide a … power into two powers of the same denomination above the second is impossible"—in the margin of a translation of, appropriately enough, Diophantus. Then he added that he had found for it an "admirable" proof, which lamentably "the margin is too narrow to contain." The challenge to find such proof was catnip to investigators until the end of the twentieth century.

Exercises VII.A.4f

1. Why did France become the center of European mathematics around 1650?
2. From $360 = 2^3\,3^2\,5$: In how many ways can 360 be written as the difference of squares?
3. To show that different factorings give different difference-of-squares expressions, draw the Quadrant I branch of the graph of $xy = k$. Place onto the graph the points (u, v) to the right of $(\sqrt{k}, \sqrt{k})$ and (s, t) to the right of (u, v).
   a) Use the lines $x + y =$ constant to show that $s + t > u + v$. The inequality guarantees that for the different factorings $k = uv$ and $k = st$,
   $$k = ([s+t]/2)^2 - ([s-t]/2)^2 \qquad \text{and} \qquad k = ([u+v]/2)^2 - ([u-v]/2)^2$$
   are two different expressions of k as difference of squares.
   b) (Calculus) Show that rightward from $x = \sqrt{k}$, $x + y$ increases with (increasing) x. That forces $s + t > u + v$.
4. Use prime factorization to show that:
   a) If d^4 divides A^2, then d^2 divides A.
   b) If $2uv$ is a square and u and v are relatively prime, then one of them has to be $2m^2$ and the other n^2, with n an odd number relatively prime to m.
   c) If a product uvw… (any finite number of factors) is a square, and the factors are pairwise relatively prime, then each of the factors has to be a square.
   d) If d^2 divides a^2, then d divides a.

5. Pascal

Blaise Pascal (1623-1662) is best known in mathematics for his pioneering work, some of it with Fermat, on probability. In that area, he showed the importance of the number triangle (section IV.B.3) that now bears his name. However, he also worked on quadratures and tangents, especially with the cycloid. That was the trouble: He flitted from topic to topic, making brilliant contributions but too soon abandoning one interest for the next (including a long retreat into religion). Chasing one interest, he built calculating machines. In the sciences, he advanced the study of fluid pressures, including Torricelli's discoveries in air pressure and the principles of hydraulics.

Our target just now is not his areas and tangents. We need instead another of his interests, using the principle of **mathematical induction** to prove statements about the natural numbers.

Imagine a sentence, which we will abbreviate by P(n), that claims something about the natural number n. A perfect example is

$$1 + 3 + 5 + \dots + (2n - 1) = n^2.$$

It says that the sum of the first n odd natural numbers is the n'th square.[This was the first relation to which the principle was applied explicitly, 80 years before Pascal used it. See **Kline**, page 272.] The idea is that you can prove P(n) to be true for all natural n by doing two things:

1. Prove that P(1) is true.

2. Assume that P(m) is true. Based on that assumption and other knowledge, prove that P(m + 1) is true.

Usually #1 (the **base case**)is an easy task. The key to #2 (the **inductive step**) is the crucial first word. You begin, not by proving something, but by *assuming* something (the **induction hypothesis**). Then you use that information to prove something related. The reason the method works is, in step #1 you prove P(1). Because step #2 says that the truth of one instance implies the truth of the next one, it follows that P(2) is true. Because that one is true, P(3) is true, ….

We will give a complicated example, in the next section, of proof by induction; try the simple sentences in the exercises.

Exercises VII.A.5

1. Prove by induction that for every natural number n:
   a) $1 + 2 + \dots + n = n(n + 1)/2$.
   b) $2^n > n$.
   c) The powers of 2 add up to 1 less than the next power:
   $2^0 + 2^1 + 2^2 + \dots + 2^n = 2^{n+1} - 1$.
2. Look at the sentence below, and judge the induction proof that follows it.
   Sentence: In a set of n horses, all of the horses are of the same color.
   Proof:
   Step 1. Clearly in a set of 1 horse, all of the (equine) members are of one color.
   Step 2. Assume that in every set of m horses, all of them are of one color. Consider now a set consisting of m + 1 horses. Among those, horses #1 through #m constitute a set of m horses. By the induction hypothesis, they are of one color. At the same time, horses #2 through #(m + 1) make up a set of m horses. Hence they all have the same color. Therefore all m + 1 horses are of the same color as horses #2 through #m. That completes the proof required in the inductive step.

6. Wallis

We move to two professional mathematicians. John Wallis (1616-1703) was a cleric, but came to a prestigious mathematical post, Savilian Professor at Oxford. Like Cavalieri, he worked with infinitesimals. Like Fermat, he brought algebra to quadratures and tangents. He would have called the process "arithmetizing" the questions; the title of his 1655 *Arithmetica Infinitorum* was an illustrative choice.

a) area under power graphs

To study the area under the graph of $y = x^k$, fixed integer $k \geq 0$, Wallis first needed a number result.

He extrapolated from cases suggesting that

$$1^k + 2^k + \ldots + n^k \quad = \quad n^{k+1}/(k+1) + \text{ lower powers of } n$$

for all natural numbers n. Recall the familiar cases

$$\begin{aligned} 1^1 + 2^1 + \ldots + n^1 &= n(n+1)/2 &&= n^2/2 + n/2, \\ 1^2 + 2^2 + \ldots + n^2 &= n(n+1)(2n+1)/6 &&= n^3/3 + n^2/2 + n/6. \end{aligned}$$

Less familiar is $k = 3$, but it happens that

$$1^3 + 2^3 + \ldots + n^3 = n^2(n+1)^2/4 = n^4/4 + n^3/2 + n^2/4.$$

We will indicate the proof by induction of the following general result.

Theorem 1. Let $k \geq 1$ be a fixed integer. Then for every natural number n,

$$n^{k+1}/(k+1) \quad < \quad 1^k + 2^k + \ldots + n^k \quad < \quad (n+1)^{k+1}/(k+1).$$

You can see that “lower powers of n” come from the binomial expansion of $(n + 1)^{k+1}$. The proof is easy *after* you know calculus; see Exercise 4.

To give evidence for the theorem, fix $k = 6$. The extension to general k will be evident, and fixing k makes it clearer that the induction is on n.

The base case (step #1) is immediate:

$$1^7/7 \ < \ 1^6 \ < \ (1+1)^7/7.$$

To do the inductive step, assume that

$$n^7/7 \quad < \quad 1^6 + 2^6 + \ldots + n^6 \quad < \quad [n+1]^7/7.$$

Then

$$n^7/7 + (n+1)^6 < \quad 1^6 + 2^6 + \ldots + n^6 + (n+1)^6 \quad < \quad [n+1]^7/7 + (n+1)^6.$$

On the right, we have

$$\begin{aligned} [n+1]^7/7 + (n+1)^6 &= \big([n+1]^7 + 7(n+1)^6\big)/7 \\ &< ([n+1]+1)^7/7 \qquad \text{(by the binomial theorem)} \\ &= (n+2)^7/7. \end{aligned}$$

On the left, we have

$$\begin{aligned} n^7/7 + (n+1)^6 &= n^7/7 + (n^6 + C_1^6 n^5 + \ldots + C_5^6 n + 1) \\ &= (n^7 + 7n^6 + 7C_1^6 n^5 + \ldots + 7C_5^6 n + 7)/7. \end{aligned}$$

There we have used the binomial (or combinatorial) coefficients $C_1^6, \ldots, C_5^6$. (We mentioned them in section VI.C.5b.) Multiply those by 7 to find

$$\begin{aligned} 7C_1^6 &= 7 \times 6/1 \quad > \quad (7{\times}6)/(1{\times}2) \quad = C_2^7. \\ &\ldots \\ 7C_5^6 &= 7 \times (6{\times}5{\times}4{\times}3{\times}2)/(1{\times}2{\times}3{\times}4{\times}5) \\ &> (7{\times}6{\times}5{\times}4{\times}3{\times}2)/(1{\times}2{\times}3{\times}4{\times}5{\times}6) \quad = C_6^7. \end{aligned}$$

(About these inequalities, see the next paragraph.) We therefore have

$$\begin{aligned} n^7/7 + (n+1)^6 &= (n^7 + 7n^6 + 7C_1^6 n^5 + 7C_2^6 n^4 \ldots + 7C_5^6 n + 7)/7 \\ &> (n^7 + 7n^6 + C_2^7 n^5 \ldots + C_6^7 n + 1)/7 \\ &= (n+1)^7/7. \end{aligned}$$

From the inductive hypothesis, we have concluded that

$$(n+1)^7/7 \quad < \quad 1^6 + 2^6 + \ldots + n^6 + (n+1)^6 \quad < \quad (n+2)^7/7.$$

That establishes the theorem for $k = 6$ by mathematical induction.

You can view those $7C_k^6$ inequalities in terms of Pascal’s triangle. They amount to the statement that the numbers on line #7 of the triangle, except for $C_1^7 = 7$ at the second spot, are less than 7 times the line #6 numbers to their upper lefts.

The lines are

$$\begin{array}{lccccccccccccccc}
\#6 & & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 & \\
\#7 & 1 & & 7 & & 21 & & 35 & & \mathbf{35} & & 21 & & 7 & & 1.
\end{array}$$

For example, the second **35** is 20 + 15, which is less than 7 × 20 because on line #6,

$$15 \;=\; (6{\times}5{\times}4{\times}3)/(1{\times}2{\times}3{\times}4) \;<\; 6\,(6{\times}5{\times}4)/(1{\times}2{\times}3) \;=\; 6 \times 20.$$

With the numerical relation in hand, Wallis gives the area under the power graph.

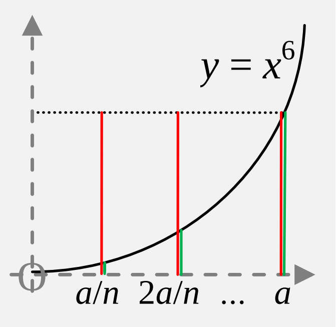
Theorem 2. The area of the region under the graph of $y = x^6$, from $x = 0$ to $x = a$, is $a^7/7$.

In the figure at right, the verticals (green) under the graph at $x = a/n, 2a/n, \ldots, na/n$ are infinitesimals of the region under the graph. Their heights are

$$(a/n)^6, (2a/n)^6, \ldots, (na/n)^6.$$

Their extensions (red) to the top of the enclosing rectangle are all a^6 tall. Therefore the ratio of the sum of the region's infinitesimals to the sum of the rectangle's infinitesimals is

$$\begin{aligned} r_n &= (1^6a^6/n^6 + 2^6a^6/n^6 + \ldots + n^6a^6/n^6) \,/\, (n\,a^6) \\ &= (1^6 + 2^6 + \ldots + n^6) \,/\, n^7. \end{aligned}$$



From Theorem 1,

$$\begin{aligned} 1/7 \;<\; r_n \;<\; & (n+1)^7/(7n^7) \\ = \;& 1/7 + C^7_1/(7n) + \ldots + C^7_6/(7n^6) + 1\,/(7n^7). \quad \text{(Verify.)} \end{aligned}$$

Now substitute $n = \infty$. (Wallis actually did that; he was the first to use "∞" to symbolize infinity, and he wrote $1/\infty = 0$.) Wallis concluded that the ratio of the region's area to the rectangle's area is 1/7, and the region has area $(1/7)aa^6$.

7. Barrow

Isaac Barrow (1630-1677) was, like Wallis, a Cambridge-educated cleric. For him, the prestigious post was the Lucasian chair at Cambridge. He was oriented toward geometry, somewhat like the Italians, but he also exercised the algebra of Fermat and Wallis. He applied the algebra to seek tangents by almost exactly the method of modern introductions to them. That is, he added a tiny change to x (what we now call Δx), causing a corresponding tiny change to y (Δy), and looked at their ratio.

Look at the graph of $y = x^{3/4}$ near the point $(a, a^{3/4})$. Add, as Fermat did, a small (infinitesimal?) change h to $x = a$. A new y-value results, which Fermat's method could not relate to h, because of the fractional power. Barrow called the new y-value $a^{3/4} + v$.

The original point satisfies

$$y^4 = x^3.$$

The new point $(a + h, a^{3/4} + v)$ has to meet the same condition. That means

$$[a^{3/4} + v]^4 \;=\; [a + h]^3.$$

Multiply out by the binomial theorem, then cancel the common a^3 to write

$$4a^{9/4}v + 6a^{6/4}v^2 + 4a^{3/4}v^3 + v^4 \;=\; 3a^2h + 3ah^2 + h^3.$$

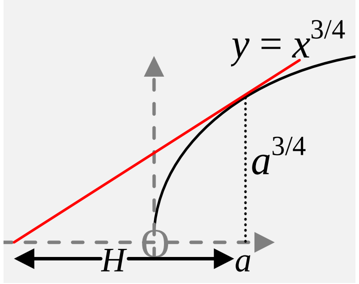
Factor v on the left and h on the right, and put their ratio as

$$v/h \;=\; (3a^2 + 3ah + h^2)/(4a^{9/4} + 6a^{6/4}v + 4a^{3/4}v^2 + v^3).$$

Barrow then said we may ignore those remaining terms with infinitesimal factors. Thus, the slope of the tangent at $(a, a^{3/4})$ is

$$3a^2/(4a^{\,9/4}) \;=\; 3/4\; a^{\,-1/4}.$$

> Being partial to geometry, Barrow insisted upon determining the tangent by means of a second point. If the tangent is not horizontal, then it is easy to locate its x-intercept. Say the intercept is H leftward from a. From the figure at right, we get
>
> $$a^{3/4}/H \;=\; \text{slope of the tangent} \;=\; 3/4\, a^{-1/4}.$$
>
> Therefore
>
> $$H = 4/3\, a,$$
>
> and another point on the tangent is $(-a/3, 0)$.



A similar argument gives the slope when the exponent is a negative fraction; see Exercise 2.

In our example, we had the y-value $a^{3/4}$ in terms of $x = a$. However, the method will produce the tangent even if the original equation does not allow us to express y “explicitly” in terms of x. It operates very much the way our “implicit differentiation” does. Thus,

$$y^3 - 7xy \;=\; x^3$$

describes *some* locus, but with an equation difficult to solve for either x or y. With Barrow’s changed values $x + h$ and $y + v$, we have

$$(y + v)^3 - 7(x + h)(y + v) \;=\; (x + h)^3,$$

and we may proceed to characterize the tangent (Exercise 3).

Exercises VII.A.7

1. We saw that the conic with vertex at (0, 0), focus at (0, f), and eccentricity ε has equation
$$x^2 \;=\; (2 + 2\varepsilon)\, fy + (\varepsilon^2 - 1)y^2.$$
a) Show that the **latus rectum** (the width of the horizontal chord through the focus) is
$$L = (2 + 2\varepsilon)\, f.$$
b) Assume that $\varepsilon < 1$, so that we are dealing with an ellipse (which might be a circle). Show that the center of the ellipse must be at $(0, a)$, with $a = f/(1 - \varepsilon)$.
c) In the ellipse from (b), show that the ratio $L/(2a)$ of the latus rectum to the long axis of the ellipse is $1 - \varepsilon^2$.
2. Apply Barrow’s method to find the slope of the tangent to $y = x^{-2/3}$ at the point $(a, a^{-2/3})$.
3. a) Apply Barrow’s method to find the slope of the tangent to the curve given by
$$y^3 - 7xy \;=\; x^3$$
at the point (2,4).
b) Characterize the tangent by finding a second point on it.
4. (Calculus) Use integration to prove that if integer $k > 0$, then for natural n,
$$n^{k+1}/(k + 1) \quad < \quad 1^k + 2^k + \ldots + n^k \quad < \quad (n + 1)^{k+1}/(k + 1).$$

Section VII.B. The Calculus

1. The State of the Art

We have arrived at about 1660. We have seen plenty of work on the problems of tangents and quadrature (of the region under a graph). Fermat must have seen, but never mentioned, the inverse relation between the answers: The slope at (x, y) of the tangent to the graph of $y = x^n$ is nx^{n-1}, and the area up to (x, y) under the graph of $y = nx^{n-1}$ is basically x^n. Boyer says that Barrow “recognized” the relation. **Struik** (page 105) states that Barrow explained it “in a difficult geometrical form,” certainly to be expected given Barrow’s preference. To develop a clear proof—it would mark what we consider to be the invention of the calculus—therefore fell to mathematicians who could integrate [no pun intended] the geometric and algebraic approaches into one infinitesimal analysis. (Those last two words sometimes serve to name the calculus.) We turn next to such men.

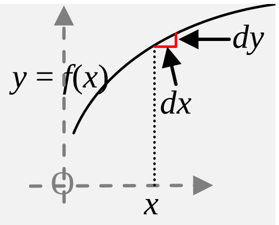
[Only the calculus carries "the," which we will sometimes skip. Nobody says "the algebra" or "the geometry." Similarly, I live in the Bronx, which nobody describes as being east of "the Manhattan" or north of "the Brooklyn."]

We said that Europe had become the center of mathematical discovery. We should note that the contributors during 120 years had been an international lot. Copernicus was Polish; Kepler German; Galileo, Cavalieri, and Torricelli Italian; Viète, Descartes, Fermat, and Pascal French; and Wallis and Barrow English. (Why were there no Spaniards or Russians in that group?)

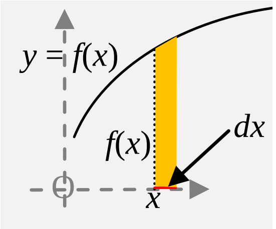
2. Leibniz

Gottfried Wilhelm Leibniz [LYPE-nits] (1646-1716), another German, contributed significantly to philosophy and logic as well as math. Like Descartes, he sought to create a general (algorithmic?) method for producing knowledge in the sciences and for deciding the validity of logical argument.

In 1672, as a government official, he traveled to Paris. There he came under the influence of the Frenchmen and Christiaan Huygens (about whom more later). The next year he traveled to London, and began to read Wallis and Barrow.



In the style of Barrow, he approached the tangent problem by adding infinitesimals he called **differentials** *dx* and *dy* (drawn at right) to *x* and *y* on the graph of $y = f(x)$. Then dy/dx would be the slope of the tangent. From this viewpoint, Leibniz chose the name **differential calculus** for the part that deals with the slope of the tangent.



Differentials also served for the area problem. In the similar lower figure, the infinitesimal of area under the graph of $y = f(x)$ is the strip (shown orange) whose height on the left (dotted) is $f(x)$ and whose width is *dx*. For the area under the graph, Leibniz summed the areas of multiple strips,

$$A \ = \ f(x_1)dx_1 + f(x_2)dx_2 + f(x_3)dx_3 + \ldots.$$

He named the process **integration** and called the sum the **integral** of the function. Those things are the subject of **integral calculus**, a name that came later. (He called it, naturally, "summation calculus.") Then he invented a notation for them,

$$\int f(x)\, dx,$$

in which the elongated "s" (for "summation") came to be called the **integral sign**. (The sign "=" for equality is likewise his invention.)

To those formulations, not in themselves new, Leibniz added some elements. For the tangent problem, he thought of the slope of the tangent to the graph of $y = f(x)$ at the point (x, y) *as itself a function*. We could substitute into dy/dx to write $df(x)/dx$, but let us use the convenient (much later) notation $f\,'(x)$. This new function $f\,'$ is called the **derivative** of *f*. For the area problem, the new wrinkle was that the formulation allowed the integrated function to be *negative*. Clearly, if $f(x_1) < 0$, then the infinitesimal of area is $[-f(x_1)]dx_1$. Therefore on intervals where *f* is negative, the integral

$$I \ = \ f(x_1)dx_1 + f(x_2)dx_2 + f(x_3)dx_3 + \ldots$$

accumulates the negative of the area *above* the graph. We will allow ourselves to continue referring to the "area below the graph," with the convention that the integral's value is the area under the part of the graph above the *x*-axis *minus* the area above the part of the graph below the axis.

[Don’t let the calculus language confuse you. You can always think in the familiar geometric terms: “derivative of” as an abbreviation for “varying slope of the tangent to the graph of”; “integral of” as abbreviating “area under the graph of.”]

a) the Fundamental Theorem

Leibniz created, in the years 1673-76, the following result:

Proposition. (The Fundamental Theorem of Calculus) The integral of the derivative of a function is the change in the function; and if we turn the integral of one function into a new function by making the right endpoint variable (the way it looks in the last figure), then the derivative of that new function is the original function.

Naturally “terms and conditions apply.” The conditions are too technical for us to detail, given our orientation. The concise statement above hides the conditions, and hides also its power.

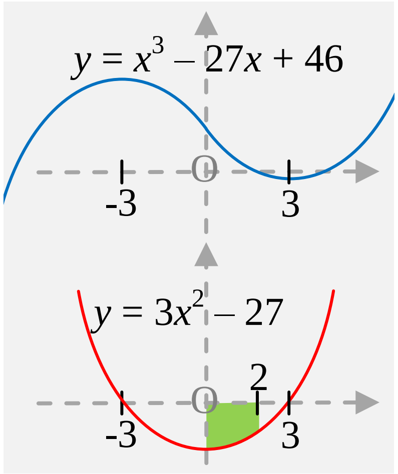
To put the statement into English—to render it in our tangent-area language—we need an example. The best kind uses a polynomial, because every polynomial meets the technical conditions. Take a polynomial we have met before, like

$$f(x) \;=\; x^3 – 27x + 46.$$

We can find the slope of its tangent term by term,

$$f\,'(x) \;=\; 3x^2 – 27(1) + 0$$

(Exercise 1). For the first part of the Theorem, f is the “function” and $f\,'$ is the “derivative.”



In the figure at right, view the graph of the derivative (red), placed below the graph of the function (blue). Areas related to the lower graph are easy to find, because it is a parabola. The region shaded green has heights (from parabola up to x-axis)

$-f\,'(0) = 27$ on the left, $-f\,'(1) = 24$ in the middle,
$-f\,'(2) = 15$ on the right.

Relying on Archimedes, we deduce (section III.A.6a) that its area is

width × weighted-average height =
$2 \times ([1/6]27 + [4/6]24 + [1/6]15) \;=\; 46.$

Therefore the integral of $f\,'(x)$ from $x = 0$ to $x = 2$ is -46 (the negative of the area above the parabola).

The message of the first half of the Fundamental Theorem is that this integral equals the change in the function. That “change” is

$$f(2) – f(0) \;=\; 0 – 46 \;=\; -46.$$

For the second half of the Theorem, turn the roles around.

Define our “original function” by

$$g(x) = 3x^2 – 27.$$

To avoid complication, we restrict ourselves to the region rightward from $x = 3$, where g is positive.

The region under the graph from $x = 3$ to a variable value x has area given by the same width-height formulation. Denote by m the value halfway between 3 and x. Then the area is

width × weighted-average height = $(x – 3)\big([1/6]0 + [4/6][3m^2 – 27] + [1/6][3x^2 – 27]\big).$

Substituting $m = (3 + x)/2$ and simplifying, we find the area to be

$A(x) \;=\; (x – 3)(x^2 + 3x – 18) \;=\; x^3 – 27x + 54.$ (Verify!)

This is the variable integral, the new function.

According to the second half of the Theorem, the derivative of $A(x)$ should be the original $g(x)$. Fermat's method confirms that

$$A'(x) \;=\; 3x^2 - 27 \;=\; g(x).$$

b) one calculus lesson

The derivative carries information about change.

Let a and $b > a$ be any two values between $x = 0$ and $x = 3$, and as above let

$$f(x) \;=\; x^3 - 27x + 46.$$

By the first half of the Theorem, the change $f(b) - f(a)$ is the integral of f' from $x = a$ to $x = b$. That integral is negative, we have noted, because f' is negative there. From

$$f(b) - f(a) \;<\; 0, \qquad \text{we get} \qquad f(b) < f(a).$$

The function has smaller values toward the right.

We conclude that *as long as the derivative is negative, the function decreases* (as x increases).

By a similar argument, after $x = 3$ the integrals of f' are positive, the changes in f are positive, the values of f get larger toward the right. In other words, *while the derivative is positive, the function increases*. (Compare Exercise 4. See also Exercise 3.)

In between, we have the place $x = 3$ where the derivative is zero. There, the function changes from decreasing to increasing: Its graph has a low point. That illustrates Fermat's theorem that maxima and minima are found among the places where the tangent has zero slope. (Compare Exercise 5.)

Exercises VII.B.2b

1. Use Fermat's method (add h to a, as in section VII.A.4d) to show that:
   a) The derivative of a sum is the sum of the derivatives. In symbols: If f and g are functions, then at the place where $x = a$, the slope of the tangent to the graph of
   $$y \;=\; f(x) + g(x)$$
   is $f'(a) + g'(a)$.
   b) The derivative of a multiple is that multiple of the derivative: The slope where $x = a$ for
   $$y = k\,f(x), \qquad k \text{ a fixed number,}$$
   is the multiple $[k\,f'(a)]$ of the slope for f.

2. Show that
   $$f(x) = x^3 - 27x + 46 \qquad \text{and} \qquad g(x) = x^3 - 27x + 54$$
   have the same derivative.

3. For what value of x beyond $x = 3$ has
   $$f(x) = x^3 - 27x + 46$$
   increased back to $f(0)$?

4. Suppose a and b are two values with $b > a \geq 3$. For
   $$f(x) = x^3 - 27x + 46:$$
   a) Show *algebraically* (not using the Theorem) that $f(b) > f(a)$ (the function increases).
   b) Show that the area under the graph of the derivative
   $$f'(x) = 3x^2 - 27,$$
   between $x = a$ and $x = b$, equals the change $f(b) - f(a)$.

5. The figure in subsection (a) shows the graph of
   $$y = x^3 - 27x + 46$$
   with a high point (maximum) in Quadrant II. Does such a maximum really exist?

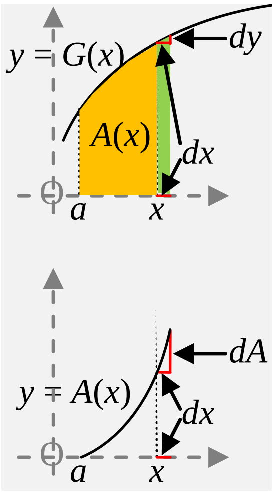
---

c) the historical import

Step back now, and see how the Theorem puts the ancient geometric problem of quadrature into the machinery of the tangent problem, where we process it using souped-up *algebra*. Leibniz—and as we will see, Newton—synthesized the geometric and algebraic approaches of their predecessors.

Indeed, the Theorem shows that the two problems “are inverses.”

Suppose we start with the graph of $y = F(x)$ and find the slopes of its tangents. Then by the Theorem, the area under the graph of the slopes is $F(x)$, almost. That is, it is the change $F(x) – F(a)$ from whatever value a we use as the left-hand border. Find first the tangent slope, then take area under the graph of the slope, and you get the function to within the fixed value $f(a)$.



In the opposite order, start with the graph of $y = G(x)$, in the top half of the figure at right. Look at area under it. Thus, let $A(x)$ denote the area (shown orange) under the graph of G from $x = a$ to an unspecified x. The graph of $A(x)$ is in the bottom half of the figure. In that half, Leibniz added the differential dx (red horizontal) to x, resulting in dA (red vertical) added to $A(x)$. The slope of the tangent to that graph is dA/dx.

But we can also see dA and dx *in the graph of* $G(x)$. Add dx to x in the top graph. The resulting increase in $A(x)$ is the green strip. It has width dx, height $G(x)$ on the left and $G(x) + dy$ on the right. Therefore

$$dA = G(x)dx + \text{some fraction of } dxdy.$$

That means

$$dA/dx = G(x) + \text{some fraction of } dy.$$

In the style of Barrow, we ignore the remaining differential. We conclude that the slope of the tangent to the graph of $y = A(x)$ is the original function $G(x)$. The “tangent to the area” gives the original function.

d) two more calculus lessons

[Indulge me.]

(i) antiderivatives

Suppose f is the derivative of the function F. We then say that F is **an antiderivative** of f.

The difference in expression—*the* derivative versus *an* antiderivative—is deliberate. The derivative of F is unique, determined by F: $f(x) = F'(x)$ is the slope of the tangent to the graph of $y = F(x)$. Antiderivatives are not unique.

We saw in Exercise 2 above that each of

$$G(x) = x^3 – 27x + 46 \qquad \textit{and} \qquad H(x) = x^3 – 27x + 54$$

is an antiderivative for $g(x) = 3x^2 – 27$. That exercise does, however, specify how antiderivatives are related: If G and H are both antiderivatives of g, then $G(x)$ and $H(x)$ must differ by a fixed number.

In that case, $G(b) – G(a)$ and $H(b) – H(a)$ have to be equal for any given a and b; by the Fundamental Theorem, each is the integral of $g(x)$ between $x = a$ and $x = b$. From

$$G(b) – G(a) \;=\; H(b) – H(a),$$

we infer

$$G(b) – H(b) = G(a) – H(a).$$

Their difference stays whatever it was at $x = a$. In the example of

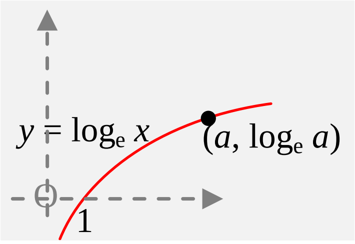
$G(x) = x^3 - 27x + 46$ and $H(x) = x^3 - 27x + 54$,
we see that $G(x) - H(x)$ has the fixed value -8.

It is implicit in the Fundamental Theorem that the problem of determining integrals amounts to that of finding antiderivatives. If F is any antiderivative of f, then the integral—the area under the graph—of f from $x = a$ to $x = b$ is $F(b) - F(a)$. (Remember that these discussions are predicated on the assumption that the functions involved satisfy those technical conditions of the theorem.)

(ii) the logarithm and exponential functions

The other lesson returns to our exceptional case. Recall our decision (section VII.A.4e(iii)) that the area under the graph of $y = 1/x$, from $x = 1$ to an $x = t$, is log t to an unknown base. In Exercise 4 of that section, we symbolized the base by e and put its value between 2 and 3.

At right, we draw the graph (red) of $y = \log_e x$. Let us examine the slope of its tangent at a general point $(a, \log_e a)$. (We will allow a to be between 0 and 1. See Exercise 2.)



In Fermat's language (mathematically, not French),

$$\text{slope} = (\log_e [a + h] - \log_e a)/h,$$

with h infinitesimally small. There, all we can do algebraically is use two properties of logarithms. They let us rewrite

$$\text{slope} = (1/h) \log_e ([a + h]/a)$$
$$= \log_e (1 + h/a)^{1/h}.$$

On the other hand, in the language of the Fundamental Theorem, the slope is the derivative of the integral of $1/x$. Therefore the slope is $1/a$. Those two unlike answers must match.

To match that last logarithm expression and $1/a$, write

$$1/a = \text{slope}$$
$$= 1/a\, [a \log_e (1 + h/a)^{1/h}]$$
$$= 1/a\, [\log_e (1 + h/a)^{a/h}]$$
$$= 1/a\, [\log_e (1 + H)^{1/H}],$$

where now H is infinitesimally small. It must be the case that

$$\log_e (1 + H)^{1/H} = 1.$$

Consequently $(1 + H)^{1/H}$ has to be e.

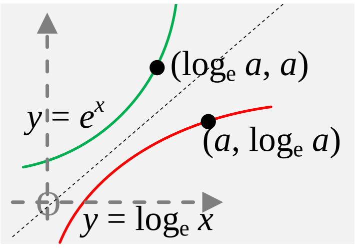
In view of their equality, we may approximate e with reasonable calculations. Thus,

$H = 1/128$ gives $e \approx (1 + 1/128)^{128} \approx 2.71$,

a calculation we can execute by repeated squaring.

Now interchange the roles of x and y. The graph of $y = \log_e x$ turns into the graph of $x = \log_e y$ (shown green at right). That equation is not how we usually describe graphs; write instead

$$y = e^x.$$



Clearly the green graph is the reflection of the red one about the 45° line (dashed black). Therefore the tangent slope at a point on the green graph is the reciprocal (not negative) of the slope at the corresponding point on the red. At $(a, \log_e a)$ on the red, the slope is $1/a$. Accordingly, at $(\log_e a, a)$ on the green, the slope is a. Call that last point (b, e^b). Then at the point (b, e^b) on the graph of $y = e^x$, the slope of the tangent is e^b. The exponential function $h(x) = e^x$ is its own derivative.

Exercises VII.B.2d

1. A bank says that it pays interest to deposit accounts at an annual rate of 1% "compounded daily." That means that every day, it adds 1/365 of 1% to your balance: It multiplies your account's balance by (1 + 0.01/365).
   a) Show without calculating that a deposit of $100 will grow in one 365-day year to approximately $\$100e^{0.01}$.
   b) Use a scientific calculator to figure out the "effective rate," meaning the total interest for the year divided by the original $100. (Banks call it the APY).
2. We have stated that the area under the graph of $y = 1/x$ from $x = 1$ to $x = a$ is $\log_e a$. That assumes $a > 1$. Does the statement remain true if $0 < a < 1$? What convention allows us to keep the statement as written?

3. Huygens

It pays to elaborate the discoveries of Christiaan Huygens (1629-1695). He was one of the most important figures in the history of science. In physics, his contributions include the wave theory of light. He used it to explain refraction—the explanation is inherent in Exercise V.A.4:1—and polarization. In astronomy, he discovered Saturn's "ring" (singular) and the big moon Titan. For our subject, he made extensive study of curves, especially (like Pascal) the cycloid. He introduced a way to determine their lengths, along with the concept of curvature (Boyer). He also published the first book (*On Reasoning in Games of Chance*) on the newly-born probability.

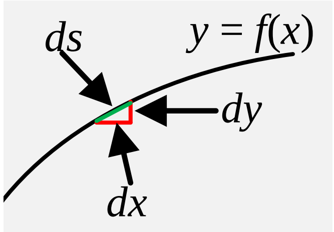
[Huygens was Dutch, but many of his accomplishments came during his years in Paris. We usually say his name HY-ggens, because English lacks the sound of the Dutch "g." It is slightly guttural. In writing, the usual way to suggest the sound is with the combination "khee." With that approximation, we would say HOOKH-yens.]

a) rectification

"Quadrature" is our Latin-based name for finding area. It reflects "squaring," the Greek idea of determining area by constructing an appropriate square. In the same way, **rectification** is our Latin-based name for finding length, reflecting the idea of "straightening" a curve to measure it. Let us study rectification in general and in two specific cases.

(i) differential of arc length

We noted that Leibniz introduced the differentials *dx* and *dy*. At right, we see them (red) in a magnified view of the figure from section VII.B.2. Maybe Huygens suggested them, but one thing that is definitely his idea is the hypotenuse *ds* (green) of their right triangle. Huygens called it the **differential of arc length**. It is an old idea that at the infinitesimal level, you cannot distinguish the curve from its chord. Therefore rectifying the curve amounts to summing the differentials of arc length. In the language of Leibniz, it comes down to integrating *ds*.



(ii) integrating the differential

From the right-triangle description, we have

$$(ds)^2 = (dx)^2 + (dy)^2.$$

[Henceforth we leave out those parentheses.]

To add up those things, it helps to rewrite

$$ds^2 \quad = \quad (1 + dy^2/dx^2)\, dx^2$$

$= \quad \surd(1 + [dy/dx]^2)\ dx.$

The last suggests that if you can write the derivative function dy/dx—the slope $f'(x)$ of the tangent to the graph of $y = f(x)$—then rectifying the graph of f is a matter of integrating the function

$\surd(1 + [f'(x)]^2)$.

Look at the most elementary example that is not already straight, and you see immediately that the task is easy to describe and hard to do.

Take $f(x) = x^2$. We know $f'(x) = 2x$. To find the length of a part of the graph of f, we must integrate

$g(x) = \surd(1 + 4x^2)$.

There is no obvious geometric way to find areas under the graph of g. The only alternative we know is to fall back upon the Fundamental Theorem: If G is an derivative of g, then the integral of g from $x = a$ to $x = b$ is $G(b) - G(a)$. Finding an antiderivative for $\surd(1 + 4x^2)$ is beyond our level.

You learn early in the study of integrals that antiderivatives are mostly hard to find. In fact, they are very much like constructions. That is, even some elementary ones are *demonstrably impossible* to give in reasonable terms. To do a nontrivial rectification, we need to pick our target with care.

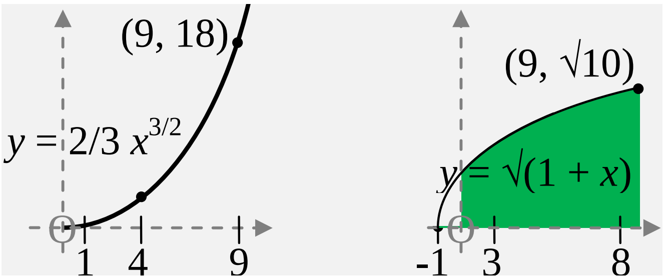
For us, finding the area under the graph of

$y \ = \ \surd(1 + [f'(x)]^2)$

is possible only if $f'(x) = \surd x$. In that case, what we have to integrate is $\surd(1 + x)$. We can do that integration in terms of area under a parabola.

From either Fermat's work on areas (section VI.A.4e) or Barrow's method for derivatives (VII.A.7), we know that the (simplest) function with that derivative is

$f(x) = 2/3\ x^{3/2}$.



In the near half of the figure at right, we draw the graph (heavy black) of

$y = 2/3\ x^{3/2}$.

[It is the upper half of the graph of

$9y^2 = 4x^3$,

which is unfortunately sometimes called the "semi-cubical parabola." Fermat actually managed a non-integrating rectification for it.]

The length of its arc from (0, 0) to (9, 18) is the integral of

$\surd(1 + [f'(x)]^2) \ = \ \surd(1 + x)$.

That integral is the area, shaded green in the right panel of the figure, under the graph of

$y = \surd(1 + x)$

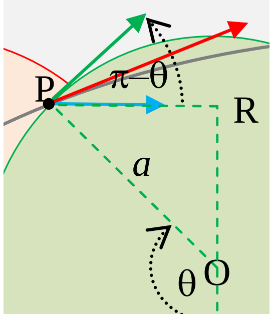
from $x = 0$ to $x = 9$. But the latter graph is just the graph of $y = \surd x$, shifted one unit to the left. Therefore the green area is simply the area under $y = \surd x$ from $x = 1$ to $x = 10$, namely

$10^{3/2}/(3/2) - 1^{3/2}/(3/2) \ = \ 2/3(10^{3/2} - 1) \ \approx \ 20.4.$

(Compare that against the estimates in Exercise 1.)

(iii) rectifying the cycloid

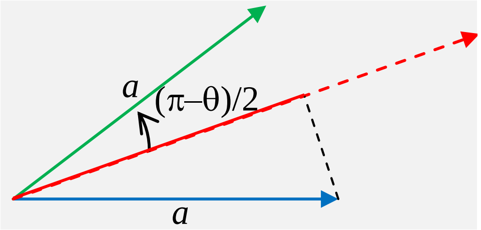
The first known rectification of the cycloid was given by Christopher Wren, architect of London's iconic St. Paul's Cathedral. Huygens later rectified the cycloid by a largely geometric method. Our approach will follow the way Roberval (section VII.A.2c, to which we will refer below) described the tangent to the cycloid, in terms of motion. We will integrate a speed. (**Boyer** says that Roberval produced a rectification ahead of Wren, but left it unpublished. Recall Roberval's secretiveness.)



At right, we reproduce the figure from the Roberval section, with slight modification. We label the angle through which the circle has rotated θ. The point tracing the cycloid (gray curve) has reached position P, and we label R the point where the horizontal through P meets the vertical through center O. At P, the tangent (along the green arrow) *to the circle* makes angle $\pi - \theta$ counter-clockwise from PR. (Verify that, by using the angles in right triangle ORP. Our figure has θ between $\pi/2$ and π, but what we say here holds before and after that interval. For example, when $\theta > \pi$, the expression $\pi - \theta$ is negative, reflecting that the green arrow points downward to either right or left.)

We observed in VII.A.2c that the point's speed (blue arrow) owing to the circle's rightward motion equals its speed (green) owing to the circle's rotation; the green and blue arrows are equally long. The equal speeds accounted for Roberval's conclusion that the tangent to the cycloid (red arrow) bisects the angle between the circle tangent and the (rightward) horizontal.

Imagine that the rolling circle rotates through 1 radian per second. Then it rolls over its circumference $2\pi a$ in 2π seconds; it has horizontal speed a (/sec). In the magnified view at right, we indicate that same speed for the tracing point to the right (blue) and along the circle tangent (green).



Therefore we have a speed a at an angle of $(\pi - \theta)/2$ on either side of the cycloid tangent, as drawn at right. Each speed contributes its projection, the solid red segment, in the cycloid tangent's direction. Those projections are

$$a \cos [(\pi - \theta)/2] \ = \ a \sin (\theta/2).$$

In the direction perpendicular to the cycloid tangent, each speed cancels the other. [That's not news; the motion of P *has to be* along the tangent to its path.] As a result—actually, as a resultant—at time t, the tracing point is covering distance along the cycloid at speed

$$2a \sin (\theta/2) \ = \ 2a \sin (t/2).$$

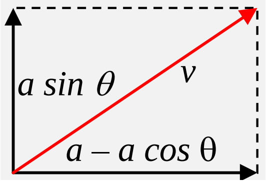
(You can derive that speed somewhat differently. Think of velocity in the more familiar way, in terms of vertical and horizontal components. The circle's rotation gives the tracing point speed a along the circle tangent. The tangent is inclined at angle $\pi - \theta$ to the horizontal. Therefore the rotation gives the point an upward speed

$$a \sin (\pi - \theta) \ = \ a \sin \theta$$

per second. (Accordingly, when θ is between π and 2π, the tracing point is going downward.) The rotation also imparts a rightward speed

$$a \cos (\pi - \theta) \ = \ -a \cos \theta.$$

At the same time, the circle's *translation* to the right adds to the point speed a rightward. The point's rightward component is therefore $[-a \cos \theta + a]$. (Hence at the ends of the arch, when $\theta = 0$ or 2π, both components of the velocity are zero; the point is stationary for an instant.) An upward component $[a \sin \theta]$ and a rightward $[a - a \cos \theta]$, drawn at right, yield a resultant speed v (length of red arrow) given by the Pythagorean theorem:



$$v = \sqrt{([a - a \cos \theta]^2 + [a \sin \theta]^2)}.$$

Now do Exercise 2.)

We have the point P tracing the cycloid with a speed

$$v = 2a \sin (t/2).$$

Huygens and Leibniz would now say that in an infinitesimal time dt, the point covers infinitesimal distance

$$ds \ = \ \text{speed} \times \text{time} \ = \ 2a \sin (t/2)\, dt.$$

To sum those differentials, we have to integrate $2a \sin (t/2)$. [If you know calculus, then by all means do the integral (Exercise 3).] Lacking an antiderivative for the sine function, we will find the total distance covered by the point the way Oresme would have three hundred years before (section V.B.3c), as the area under the graph of the speed.

In the left half of the figure below is the graph (solid curve) of

$$v = 2a \sin (t/2).$$

In the manner of Wallis (section VII.A.6a), raise the verticals at $t = 0.2°, 0.6°, 1.0°, \ldots, 359.8°$. (We have deliberately marked the abscissas by degrees, the more familiar measure. Count them; there are 900 of them. The number is not important, except for being big and convenient.) The sum of the green verticals, which reach up to the graph, is

$$2a \sin 0.1° + 2a \sin 0.3° + \ldots + 2a \sin 179.9°.$$

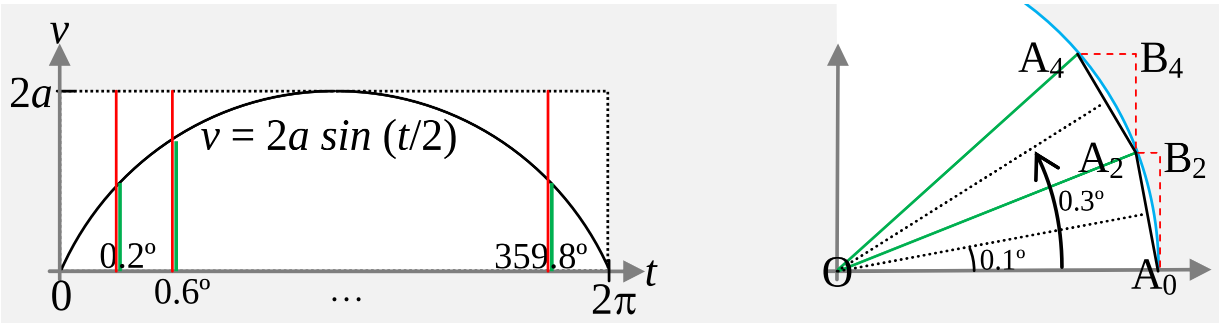
The sum of the red verticals, reaching the top of the circumscribed rectangle (white), is

$$2a + 2a + \ldots + 2a = 900\,(2a).$$

By the reasoning of Wallis, the area A of the sine arch has the fraction

$$2a\,(\sin 0.1° + \sin 0.3° + \ldots + \sin 179.9°)/(2a\,900)$$

of the area $(2\pi)2a$ of the rectangle.



To evaluate the sum of sines, look at the right panel of the figure. There, we use solid green radii to divide the Quadrant I quarter of the unit circle (blue) into 450 equal central angles of 0.2°. The first angle is A_0OA_2. Since the angle is (for us) infinitesimal, chord A_0A_2 has the same length as arc A_0A_2, namely $(\pi/2)/450$. Let B_2 be the point vertically above A_0 and horizontally right of A_2. The angle between A_0A_2 and A_0B_2 is the same as the angle between the (dotted black) bisector of angle A_0OA_2 and the horizontal, which is 0.1°. Therefore

$$A_2B_2 \ = \ A_0A_2 \sin 0.1° \ = \ (\pi/900) \sin 0.1°.$$

(Compare that argument to the Archimedes argument, section III.A.6d, for the area of the sphere.) By similar reckoning, if A_2OA_4 is the second angle, then the horizontal distance A_4B_4 from A_4 to the vertical at B_2 is

$$A_4B_4 \ = \ A_2A_4 \sin 0.3° \ = \ (\pi/900) \sin 0.3°.$$

The pattern continues until, at the top of the quarter circle, the horizontal $A_{450}B_{450}$ is

$$A_{450}B_{450} \ = \ A_{448}A_{450} \sin 89.9° \ = \ (\pi/900) \sin 89.9°.$$

All of those horizontals add up to

$$A_2B_2 + A_4B_4 + \ldots + A_{450}B_{450} \ = \ (\pi/900)\,(\sin 0.1° + \sin 0.3° + \ldots + \sin 89.9°).$$

> On the other hand, what they add up to is just the horizontal separation between A_0 at (1, 0) and A_{450} at (0, 1); it is 1. For the sum on the right,
>
> $\sin 0.1° + \sin 0.3° + \ldots + \sin 89.9° \;=\; 1/2\,(\sin 0.1° + \sin 0.3° + \ldots + \sin 179.9°)$,
>
> because the obtuse-angle sines match those of their acute supplements. Therefore
>
> $\sin 0.1° + \sin 0.3° + \ldots + \sin 179.9° \;=\; 2\,(900/\pi)\,1$.
>
> From the area proportion
>
> $A/4\pi a \;=\; (\sin 0.1° + \sin 0.3° + \ldots + \sin 179.9°)/\,900$,
>
> we conclude $A = 8a$. The length of the cycloid arch is eight radii.

There is an odd symmetry here. To sum the cycloid's infinitesimals of length, we turned the question into that of finding an area, which we evaluated by summing lengths. At the start of the chapter, we had Cavalieri seeking the area under a power graph (section VII.A.1), which problem he turned into one of summing lengths, whose sum he evaluated by turning them into areas. Sometimes in math, reversing the tool suits it better to the job at hand.

Exercises VII.B.3a

1. We found the arc length of the graph of $y = 2/3\,x^{3/2}$, from (0, 0) to (9, 18), to be about 20.4.
   a) Calculate the distance between the two points. Does it provide a lower or upper estimate of the arc length? Is it close?
   b) Calculate the length of the L-shaped path from (0, 0) to (9, 0) to (9, 18). Why is it an overestimate? Why is it so far off?

2. Show that for θ between 0 and 2π,
   $\sqrt{([a - a\cos\theta]^2 + [a\sin\theta]^2)} \;=\; 2a\sin(\theta/2)$.
   (Why is $0 \le \theta \le 2\pi$ essential?)

3. (Calculus) Evaluate $\int_0^{2\pi} 2a\,sin(t/2)dt$.

4. a) Let c and d be two angles between 0 and 90°, $c < d$. Use our quarter-unit-circle figure to argue that the area under the graph of $y = \sin x$, from $x = c$ to $x = d$, is $\cos c - \cos d$.
   b) Why does (a) imply that $-\cos x$ is one antiderivative of $\sin x$?
   c) In view of (b), what is the derivative of $f(x) = \cos x$?
   d) Do as in (a) for the graph of $y = \cos x$, use the answer to decide an antiderivative for $\cos x$, and name the derivative of $g(x) = \sin x$.

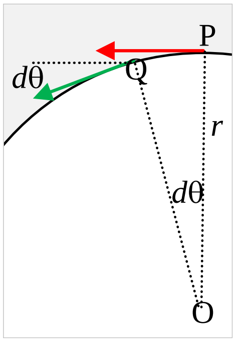
---

b) circular motion and curvature

Huygens made a discovery in mechanics, describing acceleration in circular motion, that related to the important mathematical idea of curvature.

(i) acceleration in circular motion

Huygens discovered that an object traveling at constant speed v around a circle of radius r sustains an acceleration of magnitude v^2/r in the direction of the center of the circle.



> In the figure at right, we start with the object at the top P of the circle, circling counterclockwise. Its velocity (red) is leftward with speed v. An instant (infinitesimal time) dt later, the object is at the point Q, an arc PQ = $v(dt)$ further left. The central angle POQ, labeled $d\theta$, measures
>
> $d\theta$ = arc/radius = $v(dt)/r$.
>
> At this later time, the object's velocity (green) has the direction of the circle's tangent, $d\theta$ below the horizontal. Consequently the object now has a downward

speed $v \sin d\theta$. Since $d\theta$ is infinitesimal, it equals its sine:

$$\sin d\theta = d\theta.$$

Therefore in time dt, the object has acquired a downward speed

$$v \sin d\theta \;=\; v\,(v[dt]/r).$$

That is a rate of gain per unit time—a component of acceleration toward the center—of

$$v\,(v[dt]/r)/dt \;=\; v^2/r.$$

That same green velocity gives the object a leftward speed ($v \cos d\theta$). The horizontal speed has gone from v at P to ($v \cos d\theta$) at Q, a change of ($v \cos d\theta - v$). By the half-angle formula

$$\mathit{sin}^2\,(d\theta/2) \;=\; (1 - \cos d\theta)/2,$$

we write the change in speed as

$$v(\cos d\theta - 1) \;=\; \text{-}2v\,\mathit{sin}^2\,(d\theta/2) \;=\; \text{-}2v\,(v[dt]/2r)^2.$$

Therefore the rate of gain of horizontal speed is

$$\text{-}2v\,(v[dt]/2r)^2\,/dt \;=\; \text{-}v^3 dt\,/2r^2.$$

That surviving infinitesimal is zero. [Ask Barrow.] The acceleration has zero horizontal component. We conclude that the acceleration is directed toward the center and has magnitude v^2/r.

In case these heuristics with infinitesimals give you a nagging suspicion [as they do with me and did with Bishop Berkeley (later)], here is more precise argument.

Look back at the Huygens argument. The important thing there is not simply that

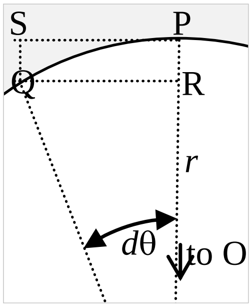
$$\sin d\theta \approx d\theta.$$

Of course they are nearly equal; they are both almost zero. The key relation is that they are *so nearly equal* that

$$[\sin d\theta]/d\theta \;=\; [\sin (v[dt]/r)]/(v[dt]/r) \;\approx\; 1.$$

It is for that reason that we write the downward acceleration as

$$\begin{aligned} \text{gain/time} \quad &= \quad v \sin d\theta/dt \\ &= \quad v\,[\sin (v[dt]/r)]/(v[dt]/r)\;\;(v[dt]/r)/[dt] \\ &= \quad v \qquad\quad 1 \qquad\qquad\qquad\qquad v/r. \end{aligned}$$



To see the relation, look at the figure at right. It has a magnified view of the top of the circle. The arc PQ has length ($r\,d\theta$) (provided θ is measured in radians) and the perpendicular QR to OP has length $r \sin d\theta$. We know that the perpendicular is a shorter path than the arc; that is,

$$r \sin d\theta \;<\; r\,d\theta, \qquad \text{and} \qquad \sin d\theta \;<\; d\theta.$$

Next, let S complete the rectangle with sides PR and RQ. The *sector* OQP of the circle is covered by triangle OQR and rectangle RPSQ. Therefore the sector's area is less than the sum of the triangle and rectangle:

$$\begin{aligned} 1/2\;r^2\,d\theta \quad &< \quad 1/2\,[\mathrm{OR}]\,\mathrm{RQ} \quad + \;[\mathrm{RP}]\,\mathrm{RQ} \\ &= 1/2\,[r \cos d\theta]\,r \sin d\theta + \;[r - r \cos d\theta]\,r \sin d\theta. \end{aligned}$$

That simplifies to

$$d\theta < \sin d\theta\,[2 - \cos d\theta]. \qquad\qquad \text{(Verify.)}$$

Our inequalities combine to give

$$1/[2 - \cos d\theta] \;<\; [\sin d\theta]/d\theta \;<\; 1.$$

That means, for example, that if $d\theta$ is within 0.01 (about 0.57°) of 0, then

$$[\sin d\theta]/d\theta \qquad \text{is between} \qquad 1 \text{ and } \;1/[2 - \cos 0.01] > 0.999\,999\,98.$$

The geometric argument just above is a standard part of calculus texts. (It relates to Exercise 1.) It is worthwhile to note that the Huygens argument did not use calculus, our choice to use differential notation notwithstanding. Its justification was the knowledge (section III.C.1a) that for a small angle, the radian measure (= arc/radius), sine, and chord/radius are indistinguishable.

(ii) curvature

At the beginning of VII.B.3b, we saw circular motion turn the tangent to a circle from horizontal to an angle below horizontal. The rate of change of that angle measures the circle's tendency to deviate from straightness—in other words, the curviness of the circle. For that reason, we call it "curvature." Huygens applied the idea to any curve, not just the circle.

In the figure back there, the tangent to the circle—the velocity of the motion—changed direction by $d\theta$ radians in time dt. The rate of change $d\theta/dt$ is not a property of the circle. It depends on how fast the object is moving. We had

$$d\theta = v\,dt/r, \qquad \text{giving} \qquad d\theta/dt = v/r.$$

For example, imagine driving at a steady 44 ft/sec (30 mph) around a circle of 100 ft radius. Then our line of sight through the windshield rotates at

$$v/r \;=\; (44\text{ ft/sec })/100\text{ ft} \;=\; 0.44\text{ (radians)/sec.}$$

That agrees with the complete turn, in which our view would rotate 2π radians in $2\pi(100)/44$ seconds. If instead we drove twice as fast, the rotation rate would likewise double.

However, what *is* a property of ("is **intrinsic** to") the circle is the rate of change of direction *per unit of distance traveled*. The infinitesimal distance ds traveled in infinitesimal time dt is ($v\,dt$), so that the change of direction per unit of distance traveled is

$$d\theta/ds \;=\; (v\,dt/r)/(v\,dt) \;=\; 1/r.$$

That rate is the **curvature** of the path—irrespective of whether the path is circular—at the given point. In view of the circle's curvature being 1/radius, Huygens called the reciprocal (1/curvature) the **radius of curvature** of the path. The idea is that around that point, the path most closely resembles a circle of that radius. [Doubtless you have heard the name in reference to roadways. Where the radius of curvature is big, we say the road curves gently, like a big circle; where it is small, we say the curve is sharp.]

Clearly the circle has the same curvature at every point. It is the only plane curve with that property. (Compare Exercise 4.) Let us look at a simple example of variable curvature.

> Recall our picture of the graph of
>
> $$y = f(x) = 1/x$$
>
> (section VII.A.4c(iii)), for which
>
> $$f\,'(x) = -1/x^2.$$
>
> The graph is curvy near (1, 1), but looks relatively straight way up or out to the right. We will approximate the curvature around both places.
>
> At (1, 1), the slope of the tangent is $f\,'(1) = -1$. Therefore the tangent has inclination
>
> $$tan^{-1}\,-1 \;=\; -\pi/4.$$
>
> At the nearby point (1.01, [roughly] 0.9901), the slope is $f\,'(1.01) = -.98030$. There the inclination is
>
> $$tan^{-1}\,-.98030 \;\approx\; -0.77545.$$
>
> The change $d\theta$ of direction is
>
> $$d\theta \;\approx\; -.77545 - (-\pi/4) \;\approx\; 0.009948.$$
>
> As for the arc length, we have no hope of integrating the differential, and no need to. The points are so close together that we simply use their distance:
>
> $$ds \;\approx\; \surd([1.01 - 1]^2 + [.9901 - 1]^2) \;\approx\; 0.01407.$$
>
> We then estimate the curvature
>
> $$d\theta/ds \;\approx\; .009948/.01407 \;\approx\; 0.707.$$

> The resemblance to √2/2 is not coincidental; the exact curvature turns out to be √2/2. Hence the radius of curvature is 2/√2 = √2. The part of the hyperbola near (1, 1) looks just like the circle through there having center at (2, 2).

We ended up with change of direction and curvature positive. If the change, and therefore the curvature, had turned out negative, we would simply have dropped the sign. A negative curvature simply indicates that the curve is turning opposite the sense in which we measure the angles. We might more accurately describe curvature as the absolute value of $d\theta/ds$.

> To the right, near the point (10, 0.1), the graph looks pretty flat. At that point, the tangent has
>
> slope $= f'(10) = -0.01$, inclination $= \mathit{tan}^{-1}\ -.01 = -0.0099997$.
>
> Nearby, at (10.01, 0.09990), it is
>
> slope $= f'(10.01) = -0.0099800$, inclination $= \mathit{tan}^{-1}\ -0.0099800 = -0.0099797$.
>
> With the distance between the points
>
> $ds = \surd([10.01 - 10]^2 + [.09990 - .1]^2) \approx 0.0100005$,
>
> we estimate the curvature as
>
> $(-0.0099797 - -0.0099997)/.0100005 \approx 0.002000$.
>
> The curve is as straight as a circle of radius 1/.002 = 500. (See also Exercise 3.)

Exercises VII.B.3b

1. Combine Fermat's method, the formula for the sine or cosine of a sum, and the discussion in (i) to show that if
   $f(x) = \sin x$ and $g(x) = \cos x$, then
   $f'(x) = \cos x$ and $g'(x) = -\sin x$.
   Compare those with the results of Exercise VII.B.3a:4.
2. (Scientific Calculator) Use the text's method to approximate the curvature of the graph of
   $y = x^2$
   at (0, 0) and (5, 25). If our usual picture of the parabola is right, then the first curvature should exceed the second.
3. (Scientific Calculator) Approximate the curvature of the graph of
   $y = 1/x$
   at (0.1, 10). The hyperbola is symmetric about the line $y = x$. This answer, then, should match the curvature at (10, 0.1), in the text.
4. Determine the curvature of the graph of
   $y = 2x + 3$,
   anywhere. Only this kind of graph has that curvature.

4. Newton

In 1669, Isaac Barrow was called, as Wallis before him had been, to be chaplain to the King. Barrow told the masters of Cambridge that they should appoint to his post his student and collaborator, whom he recognized as his intellectual superior. Thus did Isaac Newton (1642-1727) ascend at age 27 to the Lucasian chair.

a) the Fundamental Theorem

We now know that Newton discovered the theorem before Leibniz. In 1665, Cambridge came under threat of the plague. Newton spent 1665-66 home in Lincolnshire, and conceived the calculus there.

His descriptions point to his interest in the study of motion. He called x and y “fluents,” whose rates of change (“fluxions”) we denote by h and v. Then in an infinitesimal time t, x changes by ht, y by vt, and the tangent slope is $vt/ht = v/h$. (Compare those to Barrow’s changes or Leibniz’s differentials.) Similarly, the area under the graph of $y = f(x)$ increases by $f(x)ht$. Therefore the derivative of the integral (the fluxion of the area under the graph) is $f(x)ht/ht = f(x)$.

b) infinite series

Newton’s early work focused on series. He showed that you can treat some of them as if they were supersize polynomials, so that both algebra and calculus operate on them following simple rules. The best-known of his early results is the binomial series.

Proposition. (The Binomial Series) If $a > 0$ and b is between $-a$ and a, then for every real t,

$$(a + b)^t = a^t + [t/1]a^{t-1}b + [t(t-1)/1(2)]a^{t-2}b^2 + [t(t-1)(t-2)/1(2)3]a^{t-3}b^3 + \dots.$$

To put the statement into context, think of the binomial *theorem*. It gives, for example,

$$(a + b)^5 = a^5 + [5/1]a^4b + [5(4)/1(2)]a^3b^2 + [5(4)3/1(2)3]a^2b^3 + \dots.$$

In our usual writing, the sum is not unending: It stops at

$$[5(4)3(2)1/1(2)3(4)5]\, a^0b^5 = b^5.$$

But there is nothing wrong with writing it as a series. The subsequent terms

$$[5(4)3(2)1(0)/1(2)3(4)5(6)]\, a^{-1}b^6 + [5(4)3(2)1(0)(-1)/1(2)3(4)5(6)7]\, a^{-2}b^7 + \dots$$

are perfectly valid as long as $a \neq 0$.

The binomial series has the same structure. The coefficients $t/1$, $t(t-1)/1(2)$, $t(t-1)(t-2)/1(2)3$ … mimic the binomial coefficients. You should see that the exponents for which the binomial series terminates—to yield a finite sum—are precisely the nonnegative integers.

To put algebra and calculus to work on the series, let us use $(1 + x)$ to the power $t = 1/2$. (Exercise 1 has the considerably easier case $(1 - x)$ to the $t = -1$.)

The binomial series is

$$\begin{aligned}(1 + x)^{1/2} &= 1 + [1/2]x + [(1/2)(-1/2)/1(2)]x^2 + [(1/2)(-1/2)(-3/2)/1(2)3]x^3 + \dots\\ &= 1 + 1/2\, x - 1/8\, x^2 + 1/16\, x^3 - 5/128\, x^4 + 7/256\, x^5 - \dots.\end{aligned}$$

First do a slightly unconventional algebraic test: Apply the square-root algorithm to $1 + x$. We apply it as in section IV.A.2, but with no regard for place value.

In the display box at right: Each of the underlined expressions is the double of what is on the top line at the time the expression acts like a divisor; each colored term enters simultaneously on the top line and on the corresponding dashed line; and we multiply the single colored term at the top by the whole expression on the dashed line, then subtract. You can see the agreement so far. Exercise 2 takes the process two terms further.

$$\begin{array}{r|l}
 & 1 + 1/2\, x - 1/8\, x^2 + 1/16\, x^3 - \dots \\ \hline
1 & 1 + x \\
 & 1 \\ \hline
\underline{2} + 1/2\, x & \quad x \\
 & \quad x + 1/4\, x^2 \\ \hline
\underline{2 + x} - 1/8\, x^2 & \qquad -1/4\, x^2 \\
 & \qquad -1/4\, x^2 - 1/8\, x^3 + 1/64\, x^4 \\ \hline
\underline{2 + x - 1/4\, x^2} + 1/16\, x^3 & \qquad\qquad 1/8\, x^3 - 1/64\, x^4 \\
 & \qquad\qquad 1/8\, x^3 + 1/16\, x^4 - \dots \\ \hline
 & \qquad\qquad\quad -5/64\, x^4 + \dots
\end{array}$$

A more usual test is to square the series. Write out

$(1 + 1/2\, x – 1/8\, x^2 + 1/16\, x^3 – 5/128\, x^4 + 7/256\, x^5 – \ldots)^2$,

using the distributive law just as if the factors were finite sums. In the product, the constant term is 1(1) and the first-degree term is $1(1/2\, x) + 1/2\, x\, (1)$. That makes $1 + x$ so far. The next four terms are

$x^2[1(-1/8) + 1/2\,(1/2) + -1/8\,(1)] = 0$,
$x^3[1(1/16) + 1/2\,(-1/8) + -1/8\,(1/2) + 1/16\,(1)] = 0$,
$x^4[1(-5/128) + 1/2\,(1/16) + -1/8\,(-1/8) + 1/16\,(1/2) + -5/128\,(1)] = 0$,
$x^5[1(7/256) + 1/2\,(-5/128) + -1/8\,(1/16) + 1/16\,(-1/8) + -5/128\,(1/2) + 7/256\,(1)] = 0$.

[That is as far as I go. If you detect the pattern that guarantees the rest of them are zero, I will be glad to credit you with the discovery.]

Now try some calculus. Newton knew, and we have seen an illustration (Exercise VII.B.2b:1), that you can find the derivative of a sum term by term. Similarly with a series: The derivative of

$f(x) = (1 + x)^{1/2} = 1 + 1/2\, x – 1/8\, x^2 + 1/16\, x^3 – 5/128\, x^4 + 7/256\, x^5 – \ldots$

is given by

$f'(x) = 1/2 – 2/8\, x + 3/16\, x^2 – 20/128\, x^3 + 35/256\, x^4 – \ldots.$

However, we (you, in Exercise 3) can readily check that

$f'(x) = 1/2\,(1 + x)^{-1/2}$.

Those two expressions have to be equivalent. The binomial series verifies the match:

$1/2\,(1 + x)^{-1/2} = 1/2\,(1 + [-1/2]x + [(-1/2)(-3/2)/1(2)]x^2 + [(-1/2)(-3/2)(-5/2)/1(2)3]x^3 + \ldots)$.

Newton also worked on the exponential series (as well as series for sine and cosine). The exponential is of special interest.

First we think once more in terms of infinitesimals. Recall that we wrote

$e \approx (1 + H)^{1/H}$

in terms of an infinitesimal H. For reasonable x, take $H = x/1000$. From

$e \approx (1 + x/1000)^{1000/x}$, we have
$e^x \approx (1 + x/1000)^{1000}$.

By the binomial theorem,

$$\begin{aligned}(1 + x/1000)^{1000} &= 1 + [1000](x/1000) + [1000(1000–1)/1(2)]\,(x/1000)^2 \\ &\quad + [1000(1000–1)(1000–2)/1(2)3]\,(x/1000)^3 + \ldots + (x/1000)^{1000} \\ &= 1 + [1]x + [1(1–1/1000)/1(2)]\,x^2 + [1(1–1/1000)(1–2/1000)/1(2)3]\,x^3 + \ldots \\ &\quad + (x/1000)^{1000}\end{aligned}$$

Following Wallis, replace 1000 by ∞ and write

$e^x = 1 + x + x^2/1(2) + x^3/1(2)3 + x^4/1(2)3(4) + \ldots.$

(Newton, Leibniz and the others never did clarify whether an “infinitesimal” was a real number, a quantity so small that it was less than every positive real number, or some new kind of zero. They simply agreed that, say, $H/2$ is comparable to H, so that $H + H/2 = 3H/2$; but H^2 is incomparably smaller than H, so that $H + H^2 = H$. Similarly Wallis did not address the question of whether $1/H$ was actually infinite. The clarification did not come until the 1820’s.)

To work with the exponential series, try calculation first.

The series says that

$e^1 = 1 + 1 + 1/1(2) + 1/1(2)3 + 1/1(2)3(4) + \ldots.$

Take just those first five terms:

$1 + 1 + 1/1(2) + 1/1(2)3 + 1/1(2)3(4) = 65/24 \approx 2.71$,

equal to our last estimate.

Throwing away the remaining terms does not miss much. With factorial notation,

$$1/5! + 1/6! + 1/7! + \ldots \ < \ 1/120 + 1/120(6) + 1/120(6)^2 + \ldots$$
$$= 0.01.$$

(Justify all three steps. Compare Exercise 4.)

Next, check that it behaves like an exponential.

Multiplying via the distributive law, we get

$$\begin{aligned} e^s e^t &= (1 + s + s^2/2! + s^3/3! + s^4/4! + \ \ldots.)(1 + t + t^2/2! + t^3/3! + t^4/4! + \ \ldots.) \\ &= 1 + [s + t] + [s^2/2! + st + t^2/2!] + [s^3/3! + s^2t/2! + st^2/2! + t^3/3!] + \\ &\qquad [s^4/4! + s^3t/3! + s^2t^2/2!2! + st^3/3! + t^4/4!] + \ldots \\ &= 1 + [s + t] + [s + t]^2/2! + [s + t]^3/3! + [s + t]^4/4! + \ldots. \end{aligned}$$

That corresponds to

$$e^s e^t = e^{s + t}.$$

Finally, take its derivative.

From

$$\begin{aligned} g(x) &= 1 + x + x^2/2! + x^3/3! + x^4/4! + \ \ldots, \qquad \text{write} \\ g'(x) &= 1 + 2x/2! + 3x^2/3! + 4x^3/4! + \ \ldots \\ &= 1 + x + x^2/2! + x^3/3! + x^4/4! + \ \ldots. \end{aligned}$$

That accords with our earlier finding that e^x is its own derivative. (See Exercise 5.)

Exercises VII.B.4b

1. The binomial series for $(1 - x)^{-1}$, valid for $-1 < x < 1$, is

$$\begin{aligned} (1 - x)^{-1} &= 1 + -1(-x) + -1(-2)/1(2)\ (-x)^2 + -1(-2)(-3)/1(2)3\ (-x)^3 + \ldots \\ &= 1 + x + x^2 + x^3 + \ldots. \end{aligned}$$

The equality is inarguable; the last is a geometric series.
a) Do the long division of $1 - x$ into 1 to verify the series.
b) Use Fermat's method (add h to x) to show that the derivative of $(1 - x)^{-1}$ is $(1 - x)^{-2}$.
c) Write out the binomial series for $(1 - x)^{-2}$.
d) Use the "nx^{n-1}" formula to take the derivative of

$$(1 - x)^{-1} \ = \ 1 + x + x^2 + x^3 + \ldots$$

term by term, then match it to the answer in (c).

2. In the display box for the square-root algorithm, complete the last two lines shown, including the blank dashed line; then fill in the two subsequent lines, to show that the algorithm agrees with the series through the next two terms.

3. Use Barrow's method (add h to x, with the result that v is added to y) to show that the derivative of $f(x) = (1 + x)^{1/2}$ is

$$f'(x) = 1/2\ (1 + x)^{-1/2}.$$

4. a) Write the first four terms in the binomial series for

$$\sqrt{(3/4)} = (1 - 1/4)^{1/2},$$

and evaluate their sum.
b) If you stop there, then you miss the actual root by

$$[(1/2)(-1/2)(-3/2)(-5/2)/4!](-1/4)^4 + [(1/2)(-1/2)(-3/2)(-5/2)(-7/2)/5!](-1/4)^5 + \ldots.$$

Estimate how small that is.

5. a) Sketch the region under the graph of $y = e^x$ from $x = 0$ to $x = 1$, then use the sketch to estimate its area.
b) Evaluate the area. (Hint: e^x is its own derivative.)

6. Leibniz used an ingenious method (see it at Worcester Polytechnic Institute) to produce the series

   $$\sin x = x - x^3/3! + x^5/5! - x^7/7! + \dots.$$

   a) We know (Exercise VII.B.3b:1) that the derivative of sin x is cos x. Take the derivative of the sine series term by term to produce the series for cosine.
   b) From the same exercise, we know that the derivative of cos x is (-sin x). Take the derivative of the cosine series from (a) term by term to verify.
   c) Square each series, then add the squares, up to the x^6 terms. What is the sum?

c) mechanics

Newton's creation was essential in furthering his chief interest, the study of motion. Ancient mathematics had no trouble with distance covered at constant speed. Oresme and Galileo both described distance covered at constant acceleration (uniformly increasing speed). The calculus allowed analysis of *variable* rates. The power to relate quantities and their variable rates of change is what made calculus, over the centuries, the language of sciences besides mechanics, and even some social sciences.

(i) the laws of motion

Newton promulgated three principles that became the foundation of the physics of motion.

The First Law crystallized an idea that Galileo had come close to seeing, and that contradicted Aristotle. It is **inertia**, the tendency of material objects to resist *change* in motion.

Newton's First Law. An object at rest will stay at rest, and an object in motion will remain in motion …

[It is worthwhile to break up the statement of the law that way, to highlight its continuation:]
… *with constant speed along a straight line, unless acted upon by a force*.

Notice that the latter part names the agency by which change in motion comes about. The first part agrees with everyday experience. The second part does not. The followers of Aristotle certainly denied the second part, based on the experience that rolling and sliding objects will stop without continued pushing or pulling.

The Second Law describes what change a force brings about.

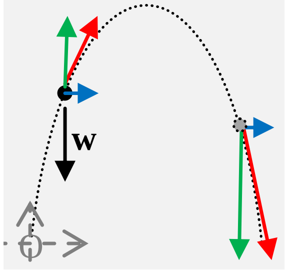
Newton's Second Law. A force acts to cause acceleration *in the direction of the force* and proportional to the magnitude of the force.

In symbols, $\mathbf{a} = k\mathbf{F}$. You see that we need to write **F** and **a** in boldface, because we have introduced the concept of vector (quantity). In the study of motion, a **vector** is a quantity whose specification requires both a magnitude and a direction. At the same time, the Law quantifies inertia, because $m = 1/k$ *measures* resistance to the effect of force. We have thus the physical attribute called **mass** and the familiar form

$$\mathbf{F} = m\mathbf{a}$$

(in which m does not need direction, is therefore "scalable," is therefore **scalar**).

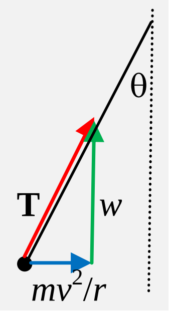
Galileo probably understood this facet of inertia. He figured that a two-pound ball accelerates downward at the same rate as a one-pounder because, subject to twice the downward pull, it also offers twice the resistance.



Remember also Galileo's cannonball tracing part of a parabola, as in the figure at right. The only force acting on the ball is its weight **w**, represented by the black arrow. That force has no part acting horizontally. Accordingly, there is no horizontal acceleration. As the ball moves along the curve, in the direction of the tangent (either of the red arrows), the rightward part (blue) of its motion is unchanging; the horizontal speed is constant. Vertically, the force points down with magnitude *w*, the (scalar) weight of the ball. Therefore on the way up, the acceleration acts against the speed. The speed decreases, at the constant rate *F*/*m* = *w*/*m*, *m* being the mass of the ball. On the way down, the acceleration acts in the direction of the speed. The speed increases, at the same constant rate. (The usual description deals with the vectors by attaching signs to the magnitudes, say + for speed or acceleration upward, - for downward. Then on the rise, speed is positive and decreasing; its rate of change is the negative number -*w*/*m*. On the fall, the speed is negative and getting more so. Therefore it is again *decreasing*. Its rate of change is again negative, the same constant -*w*/*m*.) It now becomes clear that with rates related to motion, we have the same need to combine magnitude and direction as with force. We must treat the combination of speed and direction as a vector, which gets the technical name **velocity**. In turn, that forces us to think of its rate of change, **acceleration**, in vector terms as well.

A stronger illustration comes from Huygens's discovery. Remember that an object traveling at constant speed v around a circle of radius r sustains an acceleration of magnitude v^2/r in the direction of the center of the circle. Combine that with the two Laws and you see that *keeping* an object of mass *m* circling at constant speed *requires a* **(centripetal)** *force* pulling toward the center with strength mv^2/r.

To see the force in action, tie the end of some twine or string to a reasonably small and dense load, like a key (or several taped together). When you hold the other end and let the string and key hang vertically, you feel the small tug of their weight. The key has no motion: No horizontal force acts on it, and the string's tension pulling up matches the weight pulling down. Now twirl your end gently, to make the key trace out a horizontal circle, and *focus on the string*. At any instant, the string makes some (constant) angle, labeled θ in the figure at left, with the vertical. The string is now pulling in two directions simultaneously. Part of its tension **T** is pulling upward; that component is drawn green. We know the magnitude of that pull: Just as before, it is the weight *w* of the key. It has to be; the key is not moving up or down. At the same time, part (drawn blue) of **T** is acting toward the axis of the cone the string is tracing out—in other words, toward the center of the key's circular path. That force has to be mv^2/r, where *m* is the mass of the key, *v* its speed, and *r* the radius of the circle. As the figure suggests, the magnitude *T* of **T** has increased to



$$T = \sqrt{}(w^2 + [mv^2/r]^2).$$

That greater tension is the greater tug on your fingers. Spin the key faster, and you will increase mv^2/r and feel still greater tug.

That leaves the Third Law.

Newton's Third Law. For every action, there is an equal and opposite reaction.

The statement does not sound quantitative, but it has enormous explanatory power, especially in explaining the outcome of interactions. In our age, its most important manifestation is jet and rocket propulsion. The Law dictates that the gas forced out of the open end of a jet engine *pushes back* on the engine with an equal force that propels the closed end in the opposite direction. (Jet propulsion is sometimes called "reaction" drive.) For a simpler illustration, stand facing a desk that is free to slide along the smooth floor, and give it a short, sharp push. The force you apply will likely slide the desk an inch or two ahead. The reaction force upon you will move the combination of you, floor, building, and Earth by 10^{-20} inch or so back. The reason is that the combination has around 10^{20} times the mass of the desk. If that reaction is hard to spot, try instead sitting in a chair equipped with casters, then giving the desk the same push. The reaction force will be obvious, causing you and the chair to roll back.

(ii) gravitation

From the Huygens principle and Newton's Second Law, we inferred that circular motion at constant speed implies the existence of a (possibly unseen) centripetal force related to the speed and circle's radius. The planets orbit the Sun, but not in circles nor at constant speeds. Kepler's laws say that planets orbit along ellipses, at variable distances and correspondingly variable speeds. Of necessity, some kind of force accounts for the planets' curving thus around the Sun. Analyzing such motion demanded mathematical tools with the power to handle variable distances, velocities, and accelerations. Newton had created the tools.

To make the analysis, pretend that the Sun is fixed at the origin of a coordinate system. (In reality, the Sun *must* move. By the Third Law, for whatever force it exerts on a planet, the planet tugs back equally hard. The Sun's motion is negligible to the extent that the planet's mass is negligible in comparison.) It is convenient to give an object's location—any object, planet or not—in polar coordinates. Those, too, are Newton's creation. With that setup, Newton's calculus proved that if an object obeys Kepler's Second Law—equal areas in equal times—then its acceleration is inward to or outward from the Sun. If it also obeys Kepler's First Law—like the planets, it orbits along an ellipse with focus at the Sun—then the acceleration at any time *is inward* and is inversely proportional to the square of distance at that time. Write that in the form $a = k/r^2$, with r now meaning distance from the origin and not a fixed radius. Nothing so far prevents k from being one constant for Earth, a different one for Mars, a third for Jupiter, and so on. Newton showed that Kepler's Third Law—orbital periods vary as (major axis)$^{3/2}$—implies that there is a single constant K such that every planet's acceleration has magnitude

$$a = K/r^2.$$

[Those proofs require just elementary calculus, but with skill in the rules of derivatives, polar coordinates, and vectors. If you are thus skilled and so inclined, see Appendix 2. Separately, for a wonderful account of much here, read the **Ferris** chapter on Newton, pages 103-122.]

With that planetary acceleration, Newton's Second Law implies that on a planet of mass m, the Sun exerts a force of magnitude

$$F = ma = Km/r^2.$$

Thereby, the force is proportional to the mass of the planet. By Newton's Third Law, the force is proportional to the mass M of the Sun:

$$F = ma = GMm/r^2$$

for some constant G characteristic of the solar system. That was a profound discovery. By around 1675, Newton had put a cause to the dance of the planets.

Then, Newton later told, he saw an apple fall in his mother's garden. It occurred to him that the Moon does exactly what the apple did (and what Galileo's pendulums were doing): It falls, necessarily under some force, toward Earth. We can compare the resulting accelerations of Luna and apple.

> Pretend, for approximation, that Luna orbits Earth along a circle of radius R = 30 Earths. Call the distance 240,000 miles. The implied acceleration is
>
> $$[v^2]/R = [2\pi R/\text{period}]^2/R = 4\pi^2 R/(27.3\text{ days})^2.$$
>
> With
>
> $$R = 2.4 \times 10^5\text{ mi} \times 5.28 \times 10^3\text{ ft/mi} \qquad \text{and} \qquad 27.3\text{ days} \approx 2.36 \times 10^6\text{ sec},$$
>
> we have Luna accelerating toward Earth at 9.0×10^{-3} ft/sec^2.

By Newton's time, people had a decent idea of the apple's acceleration, 32 ft/sec^2. Therefore the acceleration of the apple was

$32/(9 \times 10^{-3}) \approx 3560$

times that of Luna. The apple was (call it) 4,000 miles from the center of Earth. The inverse-square ratio (distance to apple over distance to Luna)$^{-2}$ was about

$(4000/240000)^{-2} = 3600.$

Thus, the accelerations related as the inverse-square of distance. The attraction of Earth, on the apple and on the Moon, had the same character as Sun's attraction on the planets. Newton was led to postulate:

The Law of Gravitation. Between two bodies of masses *M* and *m*, (with centers of mass) separated by distance *r*, there exists an attractive force of magnitude

$$F = G\,Mm/r^2,$$

where *G* is a *universal* constant.

Some notes are in order here.

1. The "27.3 days" figure is not a mistake. Imagine we start counting at Full Moon. In 27.3 days, Luna completes one revolution. However, it does not thereby reach the next Full Moon. During the interval, Earth goes roughly 27°, about 1/13 of its orbit, around the Sun. Therefore, to reach Full phase past Earth along the Sun-Earth line, Luna must go a further 1/13 of *its* orbit. That is why the moon—the phase cycle, Full to Full—spans the familiar 27.3(1 + 1/13) ≈ 29.4 days.

2. Determining the acceleration of falling objects, like the apple, is doable once you have reliable timers and tall structures from which to drop things. You can reasonably time a fall of 144 ft—St. Paul's in London, built during Newton's life, eventually reached more than 300 ft—taking 3 sec. From either Galileo's $s = at^2/2$ or Oresme's 144 ft/3 sec = (average speed) = 1/2(end speed), you calculate a speed gain of 32 ft/sec per second.

3. That Earth's attraction acts as though all Earth's mass were concentrated at its center was itself a fact that had to be established by integral calculus.

4. The Law allowed humans to "weigh" heavenly bodies. By around 1800, Earth's distance *r* from the Sun had been approximated with decent accuracy. That allowed calculation of Earth's speed, then Earth's acceleration *a*, just as we did with Luna. By then, Henry Cavendish had approximated *G* by exceedingly careful Earthbound experiment. When you have *r*, *a*, and *G*, from

$$a = F/m = GM/r^2,$$

you obtain the mass *M* of the Sun. Since the same law applies to Earth, we can find the mass of Earth. Indeed, we can find the mass of any celestial body that has satellites measurably far from the body. Jupiter and Saturn come immediately to mind.

(iii) differential equations

The gravitational force **F** is described by the equation

$$\mathbf{F} = G\,Mm/r^3\,(-\mathbf{r}).$$

In the context of the solar system, **r** is the **position vector** whose magnitude *r* is distance from the Sun and whose direction is from the Sun to the planet in question. The minus sign says that the force points in the opposite direction, planet to Sun. The division by r^3 sets the magnitude of **F** at $(GMm/r^3)r$; the force is proportional to inverse-square distance.

Combine that with **F** = *m***a** to write

$$\mathbf{a} = GM/r^3\,(-\mathbf{r}).$$

This equation puts acceleration, the rate of change of velocity, which is the rate of change of position, in terms of position. It was the first **differential equation**. [The name "derivative equation" would have been more informative.] It is an equation that relates a quantity (or multiple quantities) to its (their) rate(s) of change, the rates of change of those rates, and so on. Differential equations became a powerful tool in the scientific description of the world.

We can give an elementary example with another of Newton's discoveries (which, as a bonus, is outside of mechanics). **Newton's law of cooling** says that an object at a temperature different from its surroundings cools down or heats up (toward the ambient temperature) at a rate proportional to the temperature difference.

> Imagine placing a cup of 80° water into an oven at 200°. Assume the 200° is constant; any heat the oven gives to the water is replaced by the burner. Say the cooling rate is a tenth of the difference. Initially the water heats up at a rate of $(0.1/\text{min})(200° - 80°) = 12°/\text{min}$. In a short time, the cup reaches 81°. By that time, the heating has slowed to $(0.1/\text{min})(200° - 81°) = 11.9°/\text{min}$. But before then, the cup had reached 80.5°, and the rate had been …. You can see why the calculus was needed.

In general, let T be the temperature at time t. In the infinitesimal additional time dt, the change dT in temperature is given by

$$dT = (\text{heating rate}) \times \text{time} = 0.1(200 - T)\, dt.$$

In other symbols, the derivative dT/dt—the rate at which temperature changes per unit time—satisfies

$$dT/dt = 0.1(200 - T).$$

That is a differential equation. It prescribes how T changes with time. Notice that $200 > T$ for our cup. Accordingly, the rate of change is positive; the water temperature rises. If we had started with $T > 200$, the rate would have been negative, and T would have decreased. How can we describe the changing temperature as a function of time?

With the situation at hand, the Fundamental Theorem does not apply directly. If the rate of change were given in terms of t, so that the equation looked like

$$dT/dt = f\,'(t),$$

then straight integration would give us (the change in) T. Here, however, we have the rate in terms of T, not of t. We need an indirect approach.

> The direct question is to specify T in terms of t. Let us turn the question around and try to put t in terms of T. After all, we could have related the two differentials by
>
> $$1/[0.1(200 - T)]\, dT = dt.$$
>
> That form is amenable to the Theorem. It tells us that summing the differentials of time, to get the span of time needed to get from one temperature to another, is a matter of integrating.
>
> Fix a target: Ask how long it takes the water to get from 80° to 140°. We need to integrate
>
> $$g(T) = 10/(200 - T)$$
>
> from $T = 80$ to $T = 140$. In terms of areas, we need to evaluate the area under the graph of
>
> $$y = g(T) = 10/(200 - T) \qquad \text{between } T = 80 \text{ and } T = 140.$$
>
> That region has the same area as the one under
>
> $$y = 10/x \qquad \text{between } x = 60 \text{ and } x = 120 \qquad \text{(Exercise 2a).}$$
>
> We have agreed that the latter area is
>
> $$10\,[\log_e 120 - \log_e 60] = 10\,[\log_e (200 - 80) - \log_e (200 - 140)] \qquad \text{(Exercise 2b).}$$
>
> From that form, we conclude that the heating from 80° to T takes
>
> $$t = 10\,[\log_e (200 - 80) - \log_e (200 - T)] \text{ min.}$$
>
> (Compare Exercise 2c.) From there, it happens, we may express T as a function of t; see Exercise 2d.

Turning the question around led to an integral. It does not generally work, but whatever process works to solve a differential equation is sometimes called "integrating" it.

Integrating the equation

$$\mathbf{a} = GM/r^3\,(-\mathbf{r})$$

of motion under gravity is considerably harder. Naturally, Newton managed it. He showed that the possible paths, of an object subject only to an inverse-square force attracting it to a fixed body, form a family. It is the family of conic sections having one focus at the attracting body. Which of these sections the object follows, is determined by the position and velocity of the object at any chosen moment.

> The best moment to choose is when the object is closest to the attractor. At that point the conic's tangent—and therefore the velocity of the object—is perpendicular to the conic's axis (the segment from that point to the attractor).
>
> At that place, there exists a specific speed V that will cause the path to be a circle. For example, suppose the attractor is Earth, and we lift a package (hoping to make it a satellite) to a height of 200 mi. At that altitude, (call it) 4200 miles from Earth's center, a speed of $V \approx 17{,}200$ mi/hr parallel to the surface will send the package into a circular orbit of radius 4200 mi. Along that orbit, the speed will remain constant.
>
> Suppose that along the circle we later boost the speed to $v > V$. The satellite will rise to a higher point on the opposite side of Earth, then come back to the boost point. The orbit will become an ellipse of eccentricity $(v/V)^2 - 1$. The bigger v is, the more elongated the ellipse, until we choose $v = (\sqrt{2})V \approx 24{,}300$ mi/hr. Then the eccentricity reaches 1. That means the path is a parabola. The satellite never comes back. For that reason, this v is called **escape velocity** (at the 200 mi altitude). Once v exceeds $(\sqrt{2})V$, the path becomes a hyperbola. (The difference is: On the parabola, the satellite tends toward a "terminal speed" of zero, directed parallel to the axis; on a hyperbola, the terminal speed is the excess over $(\sqrt{2})V$, directed at an angle to the axis.)
>
> If the satellite is in the original circular orbit and we slow it to $v < V$, then the number $(v/V)^2 - 1$ is between 0 and -1. The orbit is again an ellipse. The negative sign says merely that the lowest point is not the start but instead the point on the opposite side of Earth. In that case, the absolute value $1 - (v/V)^2$ gives the eccentricity.
>
> In actual rocketry, the effect of lowering the low point is to put the satellite into the atmosphere, where it is either destroyed or slowed for convenient landing. If you squeezed all of Earth's mass into its center—so that the gravity were as before, but the air and planet were not in the way—then making v small would make the satellite pass at enormous speed by the center.

The remarks about speed apply to any object, including a ball we might throw or a cannonball Galileo might fire. We previously stated that the paths of those balls would be parabolas. The statement is only close to true. It is based on the assumption that either ball's acceleration is constant. Instead, acceleration varies with the force of Earth's gravity.

> If our ball reaches a maximum height of 10 ft ≈ 0.0019 mi, which is 4000.0019 mi from Earth's center, then up there acceleration is only $(4000/4000.0019)^2$ as great as at ground level. For the cannonball, if it reaches 528 ft = 0.1 mi high, then the ratio drops to $(4000/4000.1)^2$. For each ball, *the path is an ellipse*. The ellipse is indistinguishable from a parabola to the extent that say $(4000/4000.1)^2 \approx 0.99995$ is indistinguishable from 1.

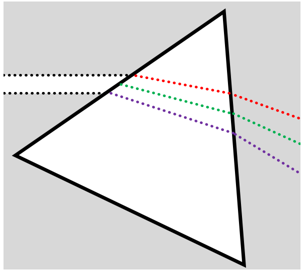
Exercises VII.B.4c

1. As you stand still on the floor, what is the reaction to your weight?

2. Sketch (roughly) the graph of $y = 10/(200 – T)$ between $T = 0$ and $T = 200$.
   a) Argue why the region under that graph from $T = 80$ to $T = 140$ is congruent to that under
   $y = 10/x$ from $x = 60$ to $x = 120$.
   b) Exactly how much is the area of the latter region in (a)? (Hint: The area under
   $y = 1/x$ from $x = 1$ to $x = b$
   is $\log_e b$.)
   c) How much time does the water in the oven take to get from 80 to 140 degrees?
   d) Solve
   $t = 10 [\log_e (200 – 80) – \log_e (200 – T)]$
   for T in terms of t. Then describe how T changes as time passes.
   e) How much time would be needed to heat the water to 199°? to 199.999°? What do you conclude about heating the water all the way to 200°?
   f) Does the answer in (e) agree with the description in (d)?
3. In a radioactive element, the atoms break down into simpler ones. The number disintegrating per unit time is proportional to how many there are. Therefore the mass m of remaining (unchanged) element decreases at a rate proportional to the mass itself:
   $dm/dt = -km$.
   In the case of radium, $k = 0.00753$/year.
   a) Solve the differential equation
   $dm/dt = -0.00753m$
   for t as a function of m, given that $m = 1$gm at time $t = 0$.
   b) How long does it take for the remaining radium to decrease to 1/2 gm? (That period is called the **half-life**. A sample of *any* size will reduce to half as much in that time.)
   c) Solve the equation from (a) for m as a function of t. What happens as time goes on?
4. By Newton’s Law, the cooling rate of a cup of coffee hotter than the surrounding air is proportional to the difference between its temperature and the ambient. Given that, why does the coffee cool faster if we pour it into a saucer?

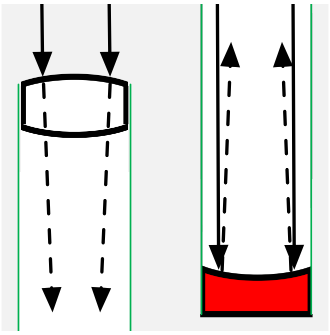
d) light

Newton made important discoveries related to light. He proposed that light consists of microscopic particles (“corpuscles”). The particle theory became a rival to the wave theory Huygens had proposed. The latter won out, because it could explain **diffraction**, the bending of light around obstacles; particles have to travel straight lines in the absence of force. Newton’s theory stayed out of favor until around 1900. Then it turned out that interaction of light with subatomic particles could only be explained by thinking of light as composed of particles (**photons**, each one a “quantum” of light energy).



More successful was his demonstration of the composition of white light. Around 1672, Newton used a slit to let a shaft of sunlight fall upon a prism, as illustrated at left. In passing through, the white light broke up into the rainbow. We noted (section V.A.4a) that at the air-to-glass entry, the red end of the rainbow is refracted less toward the normal than the violet end. At the glass-to-air exit, red is refracted less *away from* the normal. The result is a separation of the colors, as suggested in the figure. Then Newton added an experiment. He used a second slit to allow a shaft of a single color to pass through a second prism. The single color did not in turn break up into others. The single colors were the “atoms,” the indivisible constituents, of the white light.

Around 1668, Newton revolutionized telescope design. A **refracting** telescope uses a convex lens (heavy black outline in the left half of the next figure) at the top of a tube (green) to turn incident parallel rays (solid arrows) into rays (dashed arrows) converging toward a focus (out of view below the figure). Newton realized you can cause the same focusing with a concave mirror (red). There is a disadvantage: You have to suspend a smaller flat mirror along the axis of the tube (above the figure) to reflect the focused light out of the tube. However, the **reflecting** telescope has some huge advantages. The lens has two surfaces that must be accurately shaped; the mirror has just one (that has to be silvered). The lens has to be refractively uniform (equally refractive throughout); the mirror, which the light does not cross, can have variable density, flaws, even bubbles. For those reasons, reflectors are much easier and cheaper to construct than refractors of a given size. Over the three centuries after Newton, the world's biggest telescopes were always reflectors: William Herschel's 50-inch diameter (1789) in England; Mt. Wilson's 100 inches in California (1917); Hubble's 94 inches in Earth orbit; versus just 40 inches in the 1897 refractor at Yerkes Observatory (Wisconsin). (Why does diameter of lens or mirror matter?)



The lens has an additional disadvantage. Look at the figure: A convex lens vaguely resembles and acts like two prisms stuck base-to-base. Accordingly, it separates colors. That means red starlight focuses further from the lens than blue. Such **chromatic aberration** renders it impossible to examine all the light at sharp focus. The mirror avoids the aberration, because reflection does not separate the colors.

e) the astronomer

Edmond Halley (1656-1742) was a brilliant astronomer, so much that he was elected to the Royal Society and later appointed England's second Astronomer Royal. Early on he charted the southern sky from St. Helena, the distant South Atlantic island to which Napoleon would one day be exiled. His later measurements led to the discovery that the stars have **proper motion**, motion relative to other stars. [We Yanks say his name HAIL-ee. I read somewhere that the British pronunciation is HALL-ee.]

In 1684, he and others were asking whether you could explain Kepler's laws through the agency of an attraction (to the Sun) analogous to light. Light from a point source propagates out to an imaginary sphere of radius 1, then continues out to the sphere of radius r. The latter has r^2 times the area of the first. Therefore **illumination**, incident energy per unit of area, drops in proportion to $1/r^2$. The question was whether the Sun's attraction waned similarly. Halley thought to put the question to the Lucasian professor. Newton answered what we encountered in the latter part of subsection c(iii): The orbits, the paths that do not go off to infinity, are ellipses. Newton had worked it out earlier, but had to reproduce the arguments. Halley was impressed, and asked Newton to elaborate. Newton complied; he put the arguments and much background into a manuscript he showed to Halley. Halley was amazed. He begged (the reluctant) Newton to let him publish it, had it printed in 1687, paid the bill for its production. It was *Philosophiae Naturalis Principia Mathematica* (*The Mathematical Principles of Natural* [*Science*]). Halley had needed to *coax* Newton to allow publication of the most important scientific book of all time (even in a world that includes *On the Origin of Species*).

Newton was like Fermat, but considerably worse. He was so afraid of subjecting his privacy to public (or private) scrutiny that he mostly did not even communicate his discoveries to colleagues. (It seems he had no friends.) He discovered (but did not prove) the binomial theorem around 1665, but first mentioned it in a 1672 letter he sent to Leibniz via the Royal Society. The *Principia* gave the world its first look at the laws of motion and gravitation, which Newton had formulated by 1670. It gave

Newton's version of the Fundamental Theorem (from about 1665) three years after Leibniz published his version; Leibniz had discovered the Theorem some ten years after Newton.

Halley's discovery of proper motion ended forever any credibility attaching to Aristotle's idea of an immutable celestial sphere. That idea had already been dented by Tycho (mentioned in section VI.D.2a). Tycho argued that a supernova (exploding star) of 1572 had to be past the Moon and planets, because it displayed no motion and (especially) no parallax. The Moon shows parallax. If for example the northern extreme of the Moon passes across ("occults") a bright star from the viewpoint of Syracuse, then it will pass well below (south of) the star as viewed from Prague. Based on the supernova's lack of parallax, Tycho concluded that it was among the stars, even though it was transient. Later, from his own observations and reports from his network of contacts, Tycho gauged that a comet of 1577 had shown no parallax comparable to the Moon's. Accordingly, Tycho argued that comets are also travelers beyond the Moon, even though they appear and disappear.

If comets are solar system objects, Halley reasoned, then Newton's mechanics governs them. Around 1700, Halley applied Newton's laws to observations of a 1682 comet to calculate its orbit. Then he used knowledge of the masses of Jupiter and Saturn—see the note about weighing the planets at the end of subsection c(ii)—to figure how the masses of those planets would "perturb" the motion of the comet. He concluded that the comet (which he had witnessed) was the same body that had appeared in 1607 and 1531, tracing an eccentric orbit around the Sun over a period of about 76 years. Later he found records going back to ancient Chinese times, even to Babylonian times, indicating returns of the same comet. He predicted that it would return in 1757, a year he could not expect to live to see. As of December 24 that year, his prediction was unverified. The comet was spotted the next day. That Christmas Day observation, a triumph of science and in particular of Newtonian mechanics, came on the 115th anniversary of Newton's birth.

[Halley made another prediction for the future beyond him. Kepler himself had predicted that there would be transits of Venus—that Venus would cross the face of the Sun—in 1761 and 1769. Halley suggested that if astronomers mounted expeditions to scattered places on Earth, to witness the events and record the parallax of Venus, then they could calculate its distance from Earth. Thus, suppose the path of Venus across the Sun were 1/40 of Sun's size higher from Cape Town than from London. The implication would be that a (not quite north-south) distance of about 5900 mi subtends an angle of 1/40 of half a degree, or 0.00022 radian. It would imply an Earth-Venus distance of

$$(5900 \text{ mi})/.00022 \approx 27 \text{ million mi.}$$

From that one distance, you could calculate all the distances in the solar system, because Kepler's laws dictate the *relative* distances. Halley's suggestion was better on paper than in life. Read about the hardships of the expeditions British (James Cook, 1769) and French (… Le Gentil, both years) in *Sky and Telescope Magazine*'s three-part article.]

Exercises VII.B.4e

1. a) Mathematicians studied tangents and areas—the questions that lead to derivatives and integrals—before Newton. Give examples of two such people, and describe the problems they solved and how the problems relate to the calculus.
   b) In view of (a), in what sense did Newton and Leibniz "invent" calculus?

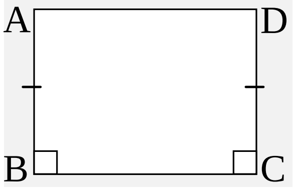
Chapter VIII. The Eighteenth Century

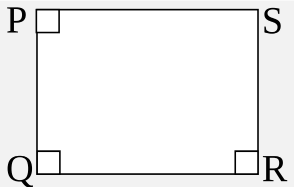
We will stretch the century slightly back, then ahead into around 1820. The biggest progress was in calculus, where the driving force was desire to perfect the analytical description of mechanics. Still, we can go back to our old way of tracking development in geometry, algebra, and number theory.

Keep in mind the political developments. England went from consolidating a world-circling empire to losing its North American colonies south of Canada. France went from the placid last years of Louis XIV to the explosion of 1789 and the years of Napoleon. Spain's European dominions shrank to the Iberian Peninsula, though she still held half of South America. Germany was not yet a country. Russia finally opened up to the rest of Europe. The great Italian cities—not Italy itself, which like Germany was not yet a state, but Venice, Florence, and the like—ceased to be powers.

Section VIII.A. Geometry

It seems like centuries since we last talked about geometry. So it was: In Euclidean geometry, nothing had happened in five hundred years. The most popular question over that time had been squaring the circle. Proving the parallel postulate may have been a close second, but the last serious work on it had been by Arabic mathematicians. However, the work picked up in the eighteenth century.

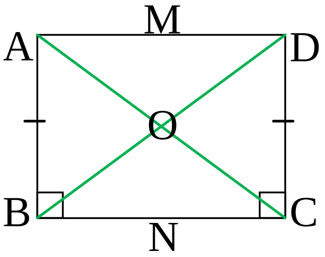




Recall what that work was about. From the time of Euclid, people had found the parallel postulate (consult section III.A.8b) unpalatable. They had tried to show that it is unnecessary, that it follows from the earlier postulates of Euclid. The Islamic world pursued a number of paths. Omar Khayyam (more familiar to us from algebra, toward the end of section V.A.3) looked at the quadrilateral at right. This **birectangular isosceles quadrilateral** has congruent perpendiculars BA and CD (the **sides**) raised at the end of segment BC (the **base**). Nasr al-Din al-Tusi (circle within a circle, section VI.D.1a) later studied the same figure. The question was whether the quadrilateral is necessarily a rectangle. Al-Hassan ibn al-Haytham (section V.A.2) looked at the **trirectangular quadrilateral**, shown at left, characterized by right angles at P, Q, and R. He faced the same question.

Once you prove some quadrilateral is a rectangle, a *long* chain of inferences establishes the parallel postulate. A good place to follow the chain is Walter Prenowitz and Meyer Jordan's *Basic Concepts of Geometry* (1965 and 1989). Roughly: The existence of one rectangle implies the existence of rectangles of all sizes; when there are rectangles of all sizes, every right triangle's angles add up to 180°; if every right triangle has that angle sum, then every triangle does likewise; and if every triangle has angle sum of 180°, then parallel lines force congruent alternate interior angles. We named that last statement the "parallel postulate," equivalent to the Euclid statement we named "Euclid's postulate" (section III.A.8b(i)).

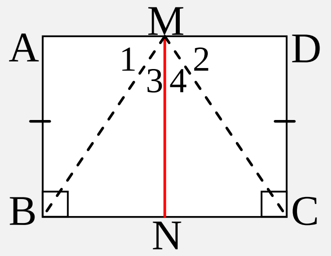
1. Saccheri



Girolamo Saccheri (Sah-CHEH-ree, 1667-1733) worked on al-Tusi's polygon, now called a **Saccheri quadrilateral**. At right, we reproduce it and add the **diagonals** AC and BD (green), which meet at O. (We will accept, as Euclid would, that they must meet within the quadrilateral.) The figure also has the midpoints M and N of the **summit** AD and the base. The segment MN joining them is the **median** of the quadrilateral. Without using the parallel postulate, Saccheri proved a series of results.

Theorems. In a Saccheri quadrilateral:
1. The diagonals are congruent.
2. The **summit angles** BAD and CDA are congruent.
3. The median is perpendicular to the base and to the summit.
4. The base is parallel to the summit.
5. The upper segments OA and OD of the diagonals are congruent, as are the lower ones OB and OC.
6. The median crosses the intersection O of the diagonals.

> Theorem 1 is Exercise 1. In view of Theorem 1, triangles CAD and BDA are congruent by SSS. Therefore angles CDA and BAD are congruent, proving Theorem 2.
>
> To prove Theorem 3, look first at triangles BAM and CDM. Because the summit angles are congruent and M is the midpoint of AD, the triangles are congruent by SAS. Therefore sides BM and CM are congruent, as are angles 1 and 2. From the congruence of the sides, we see that triangles MBN and MCN are congruent by SSS. That tells us angles 3 and 4 are congruent. Adding the angle pairs 1 and 3, 2 and 4, we conclude angles AMN and DMN are congruent. Since they are supplementary, they must be right angles. It further tells us that angles MNB and MNC are congruent. Those must likewise be right angles. We have proved Theorem 3.
>
> Theorems 4 and 5 are Exercises 2-3. For Theorem 6, ignore the median and draw the segment ON. By Theorem 5, OB is congruent to OC. Hence triangles BON and CON are congruent, by SSS. Then angles ONB and ONC are congruent, must therefore be right angles. That means ON lies along the line perpendicular to BC at N. That perpendicular is the line MN; the point O is on MN.

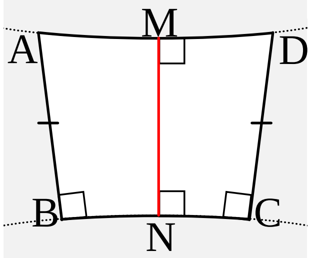


With those theorems in hand, any of a number of conclusions would establish that the quadrilateral is a rectangle.

> For one, if you could prove that the summit is congruent to the base, then triangles BCD and DAB would be congruent (SSS). That would make DAB a right angle; similarly with ADC. For another, if you could prove that the diagonals bisect, then you would know triangles BOC and DOA are congruent (SAS, via the vertical angle). That would make the summit and base congruent. Finally, if you could prove that MN is congruent to AB and CD, then you would have triangles ABN and NMA congruent. (That would be by HL, hypotenuse-leg, which does not depend on the parallel postulate.) The congruence would make BN and MA congruent; again summit and base would be congruent.

Observe that the last possibility would be an immediate consequence if you knew that parallel lines are equidistant. You cannot know that; as we stated at the end of section III.A.8b, the equidistance property is equivalent to the parallel postulate.

Only one path was left to Saccheri: Try to show that if the summit angles are either acute or obtuse, then a contradiction follows. He managed one if they are obtuse. From the assumption that they are acute, there flowed such results as summit exceeds base (AD > BC) and sides exceed median (AB = DC > MN.) Those suggested that the quadrilateral might better be rendered as at right, with the lines only seeming curved to us because of our parochial notion of straightness. Notice that this picture is faithful to known properties: The median is perpendicular to summit and base; to left and right of the median, summit and base diverge; the sides exceed the median and are still perpendicular to the base; and the summit angles are acute. However, contrary to what Saccheri evidently believed, none of what he wrote actually contradicted the earlier Euclidean postulates.



Exercises VIII.A.1. Prove, without invoking the parallel postulate, that in a Saccheri quadrilateral:

1. The diagonals are congruent.
2. The base is parallel to the summit.
3. The upper segments (OA and OD in the first figure) of the diagonals are congruent, and so are the lower segments OB and OC.

2. Lambert

Johann Heinrich Lambert (1728-1777) was Swiss. He chose to study ibn al-Haytham's trirectangular quadrilateral, needing to eliminate the possibilities that the remaining angle could be acute or obtuse.

Interestingly, he left open the "obtuse" choice. That possibility would make angle sums in triangles exceed a straight angle. To eliminate it, Saccheri had needed the assumption that lines have infinite length. (Euclid had postulated only that they could be extended.) Lambert gave the following example where things that behave like lines produce triangles with angle sums beyond 180°.

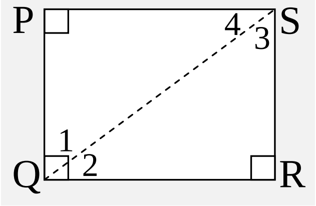
> For a creature confined to a surface—as humans were in the eighteenth century—the **geodesics** (paths of least distance) are what he has to interpret as "straight." On a sphere, the geodesics are the **great circles**, the circles with centers at the center of the sphere. We can also describe them as sections of the sphere by planes that contain the center. For humans now, they determine the great-circle routes planes nominally follow.
>
> Find a globe and follow three of them: the meridian of 75° west longitude (just west of New York) from the North Pole to the Equator (near where Ecuador, Peru, and Colombia meet) ; the Equator from the 75th meridian to the 90th (among the Galapagos Islands); and the 90th meridian north (past New Orleans) to the Pole. That is a spherical "triangle" with two 90° angles at the Equator and a 15° angle at the Pole. It has an **excess** of 15° beyond 180°. On the sphere, all "triangles" have excesses.

Lambert added a remarkable result. Slide the left edge of our triangle over to the 105th meridian, through Denver. This new triangle has an excess of 30°, and it clearly takes up twice as much Earth area. Lambert proved that if triangles have angle sums exceeding a straight angle, then *their areas are proportional to their excesses*.

He then entertained the "acute" possibility. If the quadrilateral's last angle is acute, then all quadrilaterals have angle sums under 360°. Consequently all triangles have a **defect**, a shortfall in angle sum, below 180°. Here again, Lambert showed that area is proportional to the defect.

Notice an odd thing: Under either regime—obtuse or acute—there are no similar triangles other than congruent ones. If the angles of one triangle match those of a second, then they have equal areas. In that case, the equality of size leads to matching sides.

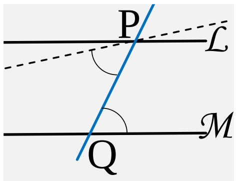
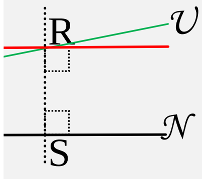
According to Boyer, Lambert was the first to recognize—certainly he was first to write explicitly—that attempts to prove the parallel postulate always end up chasing a ghost. Every such attempt had come down to establishing a reasonable statement that turned out to be equivalent to the postulate, so that the argument amounted to assuming what was to be proved. For example, in (what is now called) the **Lambert quadrilateral** at left, PS is necessarily parallel to QR. (Reason?) If we know that parallels are equidistant, then we infer that SR is congruent to PQ. Then triangles QRS and SPQ are congruent by hypotenuse-leg, angles 1 and 2 match angles 3 and 4 respectively, and (angle 3 + angle 4) is a right angle. You see that this argument hinges on the equidistance principle. We know the principle is equivalent to the parallel postulate.



3. Playfair

John Playfair (1749-1819) did not pursue the parallel postulate extensively, but he stated what is now the best-known equivalent. We will state his postulate and prove the equivalence.

Playfair's Postulate. Given a line and a point not on the line, there exists in their plane exactly one line through the given point parallel to the given line.

> Assume Playfair's postulate. At right, we have parallel lines $\mathcal{L}$ and $\mathcal{M}$ cut by a transversal (blue) at P and Q. The dashed line is constructed by producing on the left at P an interior angle congruent to the one on the right at Q. Recall that by Euclid's earlier postulates (section III.A.8b(ii)), the dashed line is parallel to $\mathcal{M}$. By Playfair's postulate, the dashed line and $\mathcal{L}$ must be one. Therefore $\mathcal{L}$ is the line that makes congruent alt-int angles. We have shown that Playfair's postulate implies the parallel (alt-int angles) postulate.
>
> 
>
>
> Conversely, assume the parallel postulate. At right, we start with point R off line $\mathcal{N}$. Drop the perpendicular (dotted) from R to S on $\mathcal{N}$, then erect the perpendicular (red) to RS at R. By the earlier postulates, the red line is parallel to $\mathcal{N}$. That gives us *one* line through R parallel to $\mathcal{N}$. Suppose now $\mathcal{U}$ (green) is a different line through point R. Then $\mathcal{U}$ is not perpendicular to RS; it makes interior angles at R unequal to the ones at S. By the parallel postulate, $\mathcal{U}$ is not parallel to $\mathcal{N}$. The red line is the *only* parallel through R. We have shown that the parallel postulate implies Playfair's.
>
> 
>

Section VIII.B. The Calculus

The progress in calculus led beyond the desired culmination of Newton's equations. Calculus and its outgrowths became indispensable for the physical description of the world, and more generally in the description of processes of change.

1. The Bernoullis

The Bernoulli family produced contributors to mathematics and physics for more than 200 years. (See the family tree in Boyer.) When the Spanish conquest of the Netherlands turned it into a dangerous place to be Protestant, much of the family left Antwerp (now Belgium, but modern Belgium was invented in 1830) for Basel (Switzerland). That was already home to Nicolaus Bernoulli and his sons, and there the boys occupied the University's Chair of Mathematics for sixty years.

a) Jacques and Jean

The sons of Nicolaus were present, effectively, at the creation. They were in contact with Leibniz right after the latter's publication of the calculus, and for years after. They advanced the subject so quickly that within twenty years they had in place much of the current form of undergraduate calculus.

Jacques Bernoulli (1654-1705) took the Basel chair in 1687. Around 1690, he persuaded Leibniz to change the name *calculus summatorius* to *calculus integralis*. For the *calculus differentialis* half, he contributed a modification to Fermat's theorem about maxima and minima (section VII.A.4d).

> Recall the animation of the cycloid (Wikipedia®). At the high points, the point tracing the curve is moving horizontally; the tangent to the cycloid is horizontal and has zero slope. At the cusps, the tracing point is moving down, stops, starts moving up. Its direction—and therefore the tangent—is vertical; the slope is undefined (or "infinite"). Jacques pointed out that the maxima and minima of a function can happen where the derivative is undefined, in addition to where it is zero (Exercise 1).

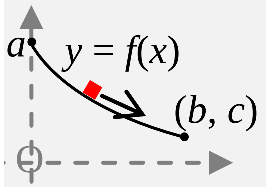
Jacques contributed as well to the study and solution of differential equations.

Jean Bernoulli (1667-1748) succeeded his brother at Basel and held the chair until his own death 43 years later. (With the Swiss, you have to track the names in multiple languages. You will find Jacques listed as Jakob or James, Jean as Johann or John.) Some of his important work came under an odd arrangement with Guillaume, marquis de L'Hôpital (1661-1704). The latter hired Jean to produce mathematical results to be published under L'Hôpital's name. The resulting publication, with some material from the author but much from Jean (including L'Hôpital's Rule, Exercise 2), became a respected and widely-used textbook.

Jean also laid the foundation for what came to be called the **calculus of variations**. The subject investigates questions of this form: Of all the functions that satisfy some condition, which one makes some integral (dependent on the function) as big or small as possible?

> The question does not have to be exotic. Remember Huygens's idea (section VII.B.3a(ii)) that the arc length of the graph of $y = f(x)$ is given by the integral of
>
> $$ds = \sqrt{(1 + [f'(x)]^2)}\, dx.$$
>
> That is an integral dependent on a function. We may ask for the function, among those whose graphs join two given points, that makes the integral smallest. That question amounts to asking for the path of least distance. We can answer *that* with no knowledge of the calculus of variations.

Look at the very first question Jean made public.



> Picture at right the graph of $y = f(x)$ *descending* from $(0, a)$ to (b, c). Imagine the red object sliding frictionlessly along the graph, pulled by gravity. The question was: Of the functions with such a graph, which will cause the object to make the trip in the minimum possible time?
>
> This is Jean's famous problem of the **brachystochrone** (from Greek for "shortest time"). It fits the form because you can express the time as an integral, much as we did for Newton's Law of Cooling (section VII.B.4c(iii)); its dependence on f is evident. You might answer with the straight line graph. It certainly gives the shortest distance. However, if you make the path start down steeply from $(0, a)$, as in the figure, then the slide accelerates more quickly. Maybe the faster speed gain will more than offset the increased distance.
>
> Jean proposed the question as a public challenge to European mathematicians. (The brothers were always posing such puzzles.) In due course, Jacques and Leibniz answered: The needed graph is the (upside-down) arch of a cycloid with cusp at $(0, a)$ and low point at (b, c). This curve had previously answered a different question. Huygens discovered that it solves the **tautochrone** ("equal time") problem: No matter where on the arc you start the object, the time the object takes to reach the bottom is the same. (Huygens had used that property to construct accurate timepieces; see Boyer.)

Just ahead of the Jacques and Leibniz solutions, a splendid one appeared in the *Philosophical Transactions of the Royal Society*. Its author, evidently publicity-shy, had requested anonymity. Jean took one look and said, "*Tanquam ex ungue leonem*," "By the claw [marks], you recognize [that it was] the Lion." In his reckoning, there existed just one Briton who could have produced so elegant an answer.

Exercises VIII.B.1a

1. a) Use Barrow's method (add h to x, triggering addition of v to y) to find the slope of the tangent to the graph of

   $$f(x) = x^{2/3}$$

   at the point $(a, a^{2/3})$.

   b) Find all a for which that derivative is either zero or undefined.

c) Use the *sign* of the derivative—remember that the sign indicates whether the graph is sloping up or down—near the places in (b) to decide whether the function has extremes (max or min) there.

2. As Fermat and the others would, we may write some of L'Hôpital's Rule as follows: If $f(0) = g(0) = 0$ and h is infinitesimal, then
$f(h)/g(h) = f'(h)/g'(h)$.
Use the Rule and what you know about the derivatives of sine and cosine to show that
$(1 - \cos h)/h^2 = 1/2$.

b) Daniel

Jean Bernoulli had three sons, of whom the elder two gained fame. Nicolaus (1695-1726) was accomplished enough to receive Peter the Great's invitation to join the Academy in Peter's new Russian capital. Nicolaus went to St. Petersburg in 1725, died there the next year. The middle son, Daniel (1700-1782), followed his brother. He remained in St. Petersburg until 1733, then returned to a succession of posts at Basel.

Daniel's strength was hydrodynamics. Among his discoveries is the fluid-flow principle (higher speed, lower pressure) that bears the family name (and explains the lift under wings). To describe fluid flow, he pioneered the field of partial differential equations (PDE's). Those equations relate functions of more than one variable and their partial derivatives. In that field, however, his best-known discovery was not about fluids. It was the **string equation**, which describes the shape of a vibrating string.

> Imagine a piano wire, fixed at both ends, vibrating (up and down, not sideways) in between. To describe its shape, we may give the (possibly negative) height y of the string above the equilibrium position, at each horizontal position x along the string as time t passes. Thus, we write y as a function
> $y = G(x, t)$
> of more than one variable.
>
> At a fixed place $x = a$ along the string, the height of the string is given by the function
> $y = G(a, t)$
> of the single variable t. The **partial derivative** of y with respect to t is the rate of change of that function per unit change in t, with x not changing from a. We can interpret it. The derivative with respect to time of any position is *velocity*. In this case, the partial derivative of y with respect to t is the vertical speed (possibly negative) of the string at the place $x = a$, at whatever time we evaluate it.
>
> At a fixed time $t = b$, the height of the string is given by the function
> $y = G(x, b)$
> of the single variable x. The rate of change of that function per unit change in x, with t not moving from $t = b$, is what we mean by the partial derivative of y with respect to x. We can interpret that one, too. When $t = b$, the wire has the shape of the graph of $y = G(x, b)$. The "partial" of y with respect to x at any given place, at this time, is the slope of the tangent to the graph at that place, at that time.
>
> Physical considerations show that there is an equation that governs all possible vibration patterns. It relates y, its partial derivatives, their partial derivatives, …. Such an equation is called a **partial differential equation**. For the string equation, Daniel produced solutions in terms of periodic functions, namely sines and cosines. Those combine to produce what we naturally call "waves."

[The invention of PDE's necessitated the creation of "ordinary differential equations" for what had been "differential equations." That is the phenomenon of **retronyms**. One familiar retronym is "film cameras," a name that had to be invented for what used to be called "cameras." (The more usual example is "acoustic guitars.")]

2. Euler

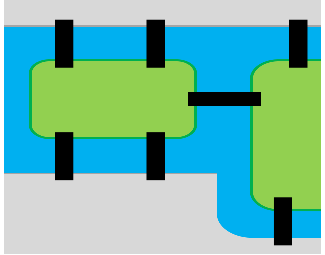
Nicolaus, then Daniel, recommended to the Academy that it import another Basel native, Daniel's best student. Like Barrow's student, this one surpassed his mentor. He surpassed everybody; he was the most prolific mathematician ever.

Leonhard Euler [OIL-er] (1707-1783) arrived in St. Petersburg a year after Nicolaus died. (Why did Euler accept, as Nicolaus and Daniel had, the invitation to such a distant place? Would Leibniz or Newton have done likewise?) He occupied the latter's post until 1741, then accepted Frederick the Great's invitation to the Berlin Academy. [You could say Euler became Mathematician Royal, but the ruler, a groupie to such luminaries as Voltaire, actually disliked the modest Euler.] By then, Euler was blind in one eye. His output continued to mark the very frontiers of mathematics. It was so extraordinary that in 1766, Catherine the Great asked him to return to St. Petersburg. [Promotion to Mathematician Imperial?] The next year, cataracts took the other eye. In his last sixteen years, the blind man *still* turned out remarkable results, dictated from a prodigious memory to servants who were not trained in math.

a) the bridges

Euler did not produce solutions so much as worlds of mathematics. He put problems he considered into contexts that inspired whole new fields of inquiry, some of which he then developed and some of which are still rich areas of study today. The best illustration is the problem of the Seven Bridges of Königsberg. It has nothing to do with calculus but is worth a detour.

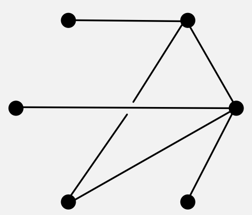
The town of Königsberg was in the Prussian province east of modern-day Poland. (Regiomontanus was born in a place of that name, but that one is in the middle of Germany. The province became East Prussia in 1920, when Germany was forced to cede the Prussian corridor along the Baltic Sea to Poland. In 1939, the Hitler government demanded it back. Poland refused, and Germany launched the invasion it would have staged regardless of the response. At World War II's end, Stalin's government kept the province and renamed the town "Kaliningrad.") A river (blue in the figure at right) runs through it, with two islands (green) in the stream. Seven bridges (black rectangles) connected the islands to each other and to the riverbanks (gray). It was a nice place to stroll around; the figure spans less than a mile. Somebody thought to ask whether it was possible to take a walk—either round trip, starting and ending at the same place, or not—that crossed each of the bridges exactly once.



Euler thought as follows. If you make a successful round trip, then with every bridge crossing, you leave a landmass you entered (or will return to at the end of the trip) via *some other bridge*. Therefore for a round trip to exist, every landmass must have leading to it an even number of bridges. That same way of counting implies that for a non-round trip, only the starting and ending places may have an odd number of connectors. Neither trip is possible in the actual setup, since all four landmasses have odd numbers of bridges.

(The argument shows that even numbers is a **necessary** condition for trip exists. That is, if even numbers is false, then trip exists is false. It does not treat the logical inverse, the question whether if even numbers is true, then trip exists is true. That would make even numbers a **sufficient** condition. It does happen to be sufficient, but further argument is required [necessary?] to prove that.)

Notice that the same reasoning applies to any number of land bodies connected by any number of bridges. Indeed, Euler recognized that it does not matter that lands and bridges are involved. There are simply places, with connections between some pairs of them. Look at the six black dots at left and the lines connecting some to others. (Any points of intersection in the middle do not count. One diagonal passes over the other; the lines connect only the big dots.) Such an arrangement is a **graph**. The dots are its **vertices**, and the lines are its **edges**. **Graph theory** is now an important area within math. It allows you to ask textbook exercises: Imagine a state road inspector, who wants to cover every inch of every road and not cover any road twice; or instead a salesman, who wants to visit every town exactly once, without regard to whether he traverses all the roads; can they manage those trips on the graph shown? But it also deals with momentous questions, owing to the modern importance of **networks**. The boards in the computer showing you this text have a staggering web of conductive paths by which various controllers must intercommunicate efficiently. You downloaded the text by means of a network of cables and transmitters connecting an array of servers working to deliver files swiftly and accurately. When the process goes wrong, you rely on the telephone network to connect you to that helpful young man in Mumbai. The design and maintenance of those webs are heavily dependent on Euler's creation.



Exercises VIII.B.2a. Here are two questions you can answer by thinking, like Euler, of unconventional ways to view and count things.

1. A tournament has 544 registered entrants. The format is **single elimination**: For each round, as many pairs of distinct players as possible are chosen at random; paired contestants play each other, and the remaining player (if any) gets a "bye"; the winners and the one with the bye advance to the next round; the process repeats until you get to two players left, who play each other in the final round to determine the champion. How many games have to be played?
2. A graph has six vertices, each connected by an edge to each of the others. Picture it as a regular hexagon with the sides and diagonals all drawn in, but some drawn in blue ink, some in red. Prove that there must exist a triangle, having vertices and edges of the graph as its vertices and sides, whose sides are of one color.
   (This question appeared in a long-ago Putnam Exam. You can extend it to a setting in the social sciences. Imagine if you had the same setup with 100 vertices. Would there necessarily exist a "ring of friends" of say 20 vertices forming a 20-gon of one color? Would there necessarily exist a "power group," a group of say 10 vertices with one color connecting each to the other 9, and every one of the remaining 90 vertices connected by that same color to at least one of the 10?)

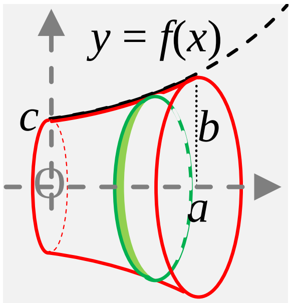
b) analysis

Euler turned the calculus into the branch of mathematics we now call "analysis." (Boyer compares Euler's synthesis of the works of Newton, Leibniz, and the Bernoullis to Euclid's codification of the geometry of predecessors like Eudoxus.) His *Introductio in Analysin Infinitorum* (1748) made "function" the central concept in analysis. It gave the first presentation of "analytic geometry" as the study of curves and surfaces entirely in terms of equations, derivatives, and integrals. He used it to extend Jean Bernoulli's study of geodesics. The book was first to treat the trigonometric functions *as* functions. It also related them to coordinates of a point on the unit circle or to ratios in a right triangle (as opposed to chords in a circle). In *Institutiones Calculi Differentialis* (*Foundations of Differential Calculus*, 1755) and *Institutiones Calculi Integralis* (three volumes ending 1770), he presented the

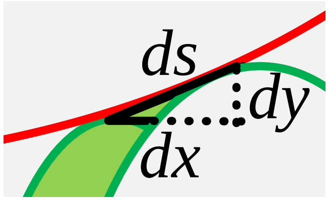
Bernoullis' and his own development of our differential and integral calculus. His contributions covered most of the theory and solution methods of our undergraduate course in differential equations. All his textbooks became immediate standards.

His work in partial differential equations extended Daniel Bernoulli's hydrodynamics and the calculus of variations. The fluid-flow PDE called "Euler's equation" has the fundamental role in fluid mechanics of Newton's **F** = *m***a** in particle mechanics. In the calculus of variations, he contributed studies of minimal surfaces.

One of those studies determined the solid of revolution of minimal area.



At right, we picture the graph (black curve) of $y = f(x)$ from $(0, c)$ to and beyond (a, b). If we revolve the region under the graph about the x-axis, it sweeps out a **solid of revolution**, outlined in red. (Compare it with Torricelli's Trumpet from Exercise VII.A.4e:5.) The question at hand is: Of the function graphs joining the two points, which one produces the solid having the least possible area?



We can express the surface area as an integral. Look at the band colored green—just the skin, not the space it encloses—in the magnified view at left. It is roughly a ring of radius $f(x)$, having therefore circumference $2\pi f(x)$. Its horizontal span is an infinitesimal dx from left to right. However, that span is not the width of the ring material. The width is the infinitesimal ds (solid black hypotenuse) of arc length along the curve. As Huygens told us,

$$ds = \sqrt{(1 + [f'(x)]^2)}\, dx.$$

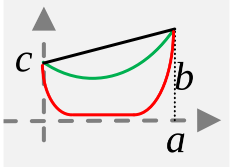
Accordingly, the band is $2\pi f(x)$ around by $\sqrt{(1 + [f'(x)]^2)}\, dx$ wide; it has surface area

$$dS = 2\pi f(x) \sqrt{(1 + [f'(x)]^2)}\, dx.$$

We conclude that the surface area of the solid is the sum of those differentials,

$$S = \int 2\pi f(x) \sqrt{1 + [f'(x)]^2}\, dx \qquad \text{from } x = 0 \text{ to } x = a.$$

The question has become one of asking which function makes some integral minimal. Treating the question falls under the calculus of variations.



Let us, ignorant of that subject, look at candidates. The straight graph (black at right) has the shortest area-sweeping length. However, it has great height. Hence it sweeps out bands of large circumference, leading to excess area. It pays to have some sag, as the green graph does. In that case, why not accept the extreme sag of the red graph, which gives the solid a long narrow "neck" of small surface area? The trouble with that one is that the long drop and rise sweep out bands of large width.

In terms of the integral: On the black graph, the factor $f(x)$ is big. It piles up excess integral. On the red graph, during both the drop and the rise, $f'(x)$ has large absolute value. Consequently the factor $\sqrt{(1 + [f'(x)]^2)}$ builds excess integral. Where do we find the middle ground?

Euler showed that the answer is a piece of the graph of

$$y = e^x + e^{-x}$$

(Exercise 1), squeezed or magnified vertically and displaced horizontally as needed. That curve had answered a different question, a challenge question issued by Brother Jacques. It described the shape of a wire or cord, hanging under its weight from supports at the two ends. Since chains, or the cables holding up a light suspension bridge, hang the same way, the curve is called a **catenary** (from Latin *catena*, chain.) Accordingly, the resulting solid of revolution is called a **catenoid**.

Exercises VIII.B.2b

1. a) Sketch the graph of $y = e^x$. (You can take it from section VII.B.2d(ii).)
   b) Flip it left for right to produce on the same set of axes the graph of $y = e^{-x}$.
   c) Use those two to sketch the graph of $y = e^x + e^{-x}$.
   d) Use Barrow's method to find the derivative of $g(x) = e^{-x}$. Does the sketch in (b) reflect that derivative?
   e) Write the derivative of $h(x) = e^x + e^{-x}$. Where is $h'(x)$ either zero or undefined? Does the sketch in (c) reflect this information?

c) infinite series

Newton's facility in treating series like finite sums became a weapon in the hands of Euler. Actually, it became something of a loose cannon. We will see later an especially imaginative—"illegal" would be a more accurate word—use he made of series. Here we look at his combination of the exponential, sine, and cosine series.

Recall that the three series are (Section VII.B.4b and Exercise 6 there)

$e^x = 1 + x + x^2/2! + x^3/3! + x^4/4! + \ldots,$ (The notation "*e*" is Euler's idea.)

$\sin x = x - x^3/3! + x^5/5! - x^7/7! + \ldots,$

$\cos x = 1 - x^2/2! + x^4/4! - x^6/6! + \ldots.$ (The notations "sin." and "cos." are Euler's.)

(Hereafter, we will refer to expressions like these as **power series**.) Notice that the terms of the exponential are split between the other two, albeit supplied with alternating signs. Euler introduced the needed signs by substituting the imaginary complex number *ix*. (The notation "*i*" for √-1 was Euler's.) Thus,

$$e^{ix} = 1 + ix + i^2x^2/2! + i^3x^3/3! + i^4x^4/4! + \ldots.$$

The powers of *i* are

$$i^1 = i, \qquad i^2 = -1, \qquad i^3 = ii^2 = -i, \qquad i^4 = i(-i) = 1, \qquad i^5 = i, \ldots.$$

Substituting them we have

$$\begin{aligned} e^{ix} &= 1 + ix - x^2/2! - ix^3/3! + x^4/4! + ix^5/5! - \ldots \\ &= (1 - x^2/2! + x^4/4! - x^6/6! + \ldots) + i(x - x^3/3! + x^5/5! - x^7/7! + \ldots) \\ &= \cos x + i \sin x. \end{aligned}$$

(In Euler's time, there were still objections to imaginary *numbers*. Imagine [no pun] resisters' reaction to imaginary *powers*.)

The equation looks a little strange, but it bears many gifts. One is the theorem named after Abraham de Moivre (1667-1754). De Moivre wrote that

$$(\cos x + i \sin x)^n = \cos nx + i \sin nx.$$

The equation follows immediately from Euler's, since the right side is

$$e^{i(nx)} = (e^{ix})^n$$

(but do the separate proof in Exercise 2).

In turn, from de Moivre's theorem, we get the multiple-angle formulas that Viète created and enlisted as aides in solving (polynomial) equations (section VI.C.5b).

First, combine the theorem with the binomial theorem and evaluate the powers of *i* to write

$$\begin{aligned} \cos nx + i \sin nx &= (\cos x + i \sin x)^n \\ &= cos^n x + \binom{n}{1} i\, cos^{n-1} x \sin x - \binom{n}{2} cos^{n-2} x\, sin^2 x - \binom{n}{3} i\, cos^{n-3} x\, sin^3 x + \ldots. \end{aligned}$$

(That notation for the binomial coefficients is Euler's, except for an underbar: $\binom{n}{1}$.)

Next, consider that a complex-number equation encapsulates two equalities: The real parts have to match, and so do the imaginary parts. Separating on the right the terms that have factor i from the others, we conclude

$$\cos nx = \cos^n x - \binom{n}{2}\cos^{n-2} x \sin^2 x + \binom{n}{4}\cos^{n-4} x \sin^4 x - \ldots,$$
$$\sin nx = \binom{n}{1} \cos^{n-1} x \sin x - \binom{n}{3}\cos^{n-3} x \sin^3 x + \binom{n}{5}\cos^{n-5} x \sin^5 x + \ldots.$$

One sum ends at $\pm\sin^n x$, the other at $(\pm n \cos x \sin^{n-1} x)$, depending on whether n is odd or even. (See a small example in Exercise 1.)

The exponential equation also resolved issues related to the restricted domains of some functions of real numbers. Consider $f(x) = \sqrt{x}$, which is undefined if x is negative. If you allow complex values, then the restriction disappears, though at a cost. We may write $\sqrt{-1} = \pm i$. We gain extension of the domain of the square-root function. The cost is the function-ness: We end up instead with a **relation**, having two values and giving us no reason to choose one over the other. Euler, who was first to identify $\log_e x$ (which he denoted by l. x [ell x]) as *an exponent* (to which you raise e to get x), extended the logarithm function to negative x.

Thus,

$$e^{i\pi} = \cos \pi + i \sin \pi = -1$$

allows us to write

$$\log_e (-1) = i\pi.$$

Again, unique value is lost: We also have

$$-1 = e^{-i\pi} = e^{\pm 3i\pi} = e^{\pm 5i\pi} = e^{\pm 7i\pi} = \ldots,$$

which means $\log_e (-1)$ has an infinity of values.

The technique allows us to define complex powers of complex numbers. We have

$$e^{i\pi/2} = \cos \pi/2 + i \sin \pi/2 = i.$$

Therefore one value of $i^{(2+i)}$ is the real number

$$(e^{i\pi/2})^{(2+i)} = e^{(i\pi/2)(2+i)} = e^{i\pi + ii\pi/2} = e^{i\pi} e^{-\pi/2} = -e^{-\pi/2}.$$

You will usually see $e^{i\pi} = -1$ written as

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

In this form, many consider it the most beautiful equation in mathematics. It displays the fundamental operations of addition and multiplication; the two identities 0 and 1 (definition later); the two most important real constants e and π (the notation π was not Euler's, but its widespread use follows his example); and the quantity i that is the gateway to the complex numbers.

Exercises VIII.B.2c

1. a) Multiply out de Moivre's relation

$$(\cos x + i \sin x)^3 = \cos 3x + i \sin 3x$$

to show that

$$\cos 3x = \cos^3 x - 3 \cos x \sin^2 x,$$
$$\sin 3x = 3 \cos^2 x \sin x - \sin^3 x.$$

b) Is the cos $3x$ formula in (a) equivalent to our old

$$\cos 3x = 4\cos^3 x - 3\cos x?$$

c) Multiply out

$$(\cos x + i \sin x)^4 = \cos 4x + i \sin 4x$$

to write "quadruple-angle formulas" for cos $4x$ and sin $4x$. Check the former against Exercise VI.C.4b:4a.

2. Write a proof by induction of de Moivre’s theorem. (Would induction have been available to de Moivre?)
3. Find values for: a) $\sqrt{i}$ b) cos i. (Hint: Write an expression for e^{-ix}.)

3. D’Alembert

Despite the Swiss, France remained the center of European mathematics. It produced an incredible line of scientists and mathematicians. In connection with the calculus, we are going to restrict our attention to just two of them. The first is Jean (le Rond) d’Alembert [dalom-BEAR] (1717-1783).

[In the eighteenth century, British development of mathematics fell considerably behind the rest of Europe. I have often heard this lack of progress blamed on the clumsiness of Newton’s method of fluxions, compared to the more dynamic method—and flexible notation—of Leibniz and the other continentals. **Boyer** is the first place where I saw the charge disputed.]

D’Alembert was already famous by 1750. Around then, he began to collaborate with Denis Diderot on the latter’s *Encyclopédie*. For more than twenty years, he was what we might call the encyclopedia’s “science editor.” His 1754 appointment as secretary (*secrétaire perpetuel*) to the *Académie des Sciences* made him the chief judge of Europe’s scientific and mathematical work. [The encyclopedia was a compendium of philosophy as well as knowledge. It was more the embodiment of the spirit of the Enlightenment than its product. As such, it was part of the opposition to monarchy. With that outlook, d’Alembert and his friends Diderot and Voltaire were among the forebears of the French Revolution.]

a) limits

D’Alembert had an early interest in hydrodynamics, and published a book on the dynamics of solids (as opposed to particles). The interest in fluids necessarily led to PDE’s, in particular to the string equation. He extended Daniel Bernoulli’s work, producing an elegantly simple form for the solutions of the equation. However, our main interest in his mathematical work is his writing on limits.

Before d’Alembert, the Irish bishop George Berkeley (1685-1753) criticized the logic of calculus, much as Zeno had done with geometry two thousand years before (section III.A.4c). Specifically, Berkeley attacked the reliance on infinitesimals (Fermat, Wallis, Barrow) and fluxions (Newton).

> Treat an example in the manner of Fermat. Subtract $f(x) = x^3$ from the nearby $f(x + h) = (x + h)^3$, then divide by h:
> $$[f(x + h) - f(x)]/h \;=\; [(x + h)^3 - x^3]/h \;=\; 3x^2 + 3xh + h^2.$$ (Check the algebra.)
> At that point Fermat set $h = 0$ to find the slope of the tangent to the graph of $y = f(x)$.

Berkeley said this was sophistry, not science. It might as well be a leap of faith. First, you assume that h is nonzero. You have to, otherwise you cannot do the division. Then you assume it *is* zero. All you could offer in defense against Berkeley’s indictment is that the method produced valid and useful results. He said the results came about because some errors cancelled others.

(Bishop Berkeley was trained in science as well as divinity. Still, it was an incident involving faith that triggered publication of his criticism; read it in **Boyer**, pages 469-470.)

D’Alembert voiced the same objections. For him, a quantity was zero, or was not zero. More important, though, he gave a way to avoid the logical problem. He said that what was necessary was to see the *limit* of the quotient. In our example, he would name $3x^2$ as the limit of

$$[f(x + h) - f(x)]/h \;=\; 3x^2 + h(3x + h)$$

because the quotient can “approach [the stated limit] nearer than by any given quantity” (Boyer). Here it is clear that forcing the quotient near to $3x^2$ is a matter of making h small. (Look also at the calculation we made for the limit of [sin θ]/θ at the end of section VII.B.3b(i), and see Exercise 1 here.)

Newton always worked with his fluents and fluxions, but it happens that at one place he anticipated d'Alembert's limit language. He wrote (**Struik**, page 111) that our quotients (the "prime ratios") "approach nearer than by any given difference" to the limits ("ultimate ratios").

b) infinity

Having no need for infinitesimals, d'Alembert could avoid the ("actually") infinite. He gave a description very much like the modern: A quantity is **infinite** if it is larger than any given number. Thus, he would agree that the ("sum" of) the harmonic series

$$1/1 + 1/2 + 1/3 + 1/4 + \ldots$$

is infinite, because it exceeds for example 10^9. (That was the example in Exercise V.B.3:4.) In contrast,

$$1/1^2 + 1/2^2 + 1/3^2 + 1/4^2 + \ldots$$

is finite. Jacques Bernoulli, who had rediscovered Oresme's argument (section V.B.3b) for the harmonic series, showed that the sum never even reaches 2. (His argument is suggested in Exercise 2.)

Exercises VIII.B.3

1. What values of r, between 0 and 1, will guarantee that

   $$1/(1 + r + r^2 + r^3)$$

   "approaches" 1/4 "nearer than by" 0.000001 (guarantee the fraction is within 10^{-6} of 1/4)? (In section VII.A.4e(i), we used Fermat's method to approximate the area under the graph of $y = x^3$. The approximation used the relation

   $$(1 - r)(1 + r^4 + r^8 + \ldots) = 1/(1 + r + r^2 + r^3)$$

   with $r = 0.99$. Fermat substituted $r = 1$ on the right to get 1/4. D'Alembert would have outlawed the substitution, because it is illegal on the left side. He would have called for the *limit* of the two sides. This exercise asks for numerical evidence that the limit is 1/4.)

2. Show that

   $$1/1^2 + 1/2^2 + 1/3^2 + \ldots + 1/1000^2 < 1 + 1 - 1/1000.$$

   (Hint: Start by showing that for integers $k \geq 2$,

   $$1/k^2 < 1/([k - 1]k) = 1/[k - 1] - 1/k.)$$

4. Lagrange

Joseph-Louis, comte de Lagrange [roughly la-GRONSH] (1736-1813) was a giant almost on the level of Euler. The scope of his contributions is vast, but they are largely too advanced for our treatment. The same was true with Euler, and will be with Gauss. For each of the three, we will touch upon those discoveries that we can describe in elementary terms. In this section, the subject is Lagrange's analysis.

At age nineteen, he was teaching artillerymen in the Military Academy at Torino. (He was actually born there, Turin, to the name Giuseppe Luigi Lagrangia.) In 1766, Frederick the Great invited him to head the Berlin Academy, at the recommendation of d'Alembert and the departing Euler. When Frederick died in 1786, Lagrange accepted Louis XVI's invitation to Paris

[Moving to Paris three years before 1789 was an interesting choice. During the Terror, it took Lagrange's fame to prevent his expulsion from France. His life was not in danger: He wasn't a French aristocrat; the "count" title would come from Napoleon in 1808. Still, he needed help. The most important help was the intercession of Antoine Lavoisier, whose genius and fame didn't do *his* head any good. Read at the American Chemical Society.]

You can get some idea of the reach of Lagrange's mathematics from things named for him. In the calculus of variations—a subject whose name he created around 1760—he discovered the fundamental partial-derivative relation now called the *Euler-Lagrange equation*. Already by 1755, he had begun to give the whole subject an elegant analytical form. He wrote about it to Euler, who had made similar discoveries. (Euler held back publishing his version. There are those, including **Boyer**, who ascribe Euler's choice—ceding priority of publication to Lagrange—to the old guy's generosity.) Lagrange then applied his variational methods to a principle in mechanics to develop *Lagrange's equations*. [Read about that "principle of least action" from Richard Feynman, no less.] Those represented a powerful refinement and extension of Newton's equations of motion. They led to the discovery of *Lagrangian points* (described later) in planetary dynamics. Separately, he turned the variational ideas back to mathematics, and developed the method of *Lagrange multipliers* for a class of "constrained optimization" problems. (See Exercise 1 for an example.)

a) series

Lagrange used infinite series to propose an analytical basis for the calculus. Nowadays, one sense of "analytical" is "having to do with calculus." For Lagrange, the word implied—as in the calculus of variations—a treatment using algebraic techniques, without the geometry needed for the approaches of Leibniz, Newton, and even Euler. (**Struik** page 134 says that in the preface to *Mécanique Analytique*, Lagrange specifically announced that there are no figures in the book, only algebraic operations.)

> Begin with any power series
>
> $$f(x) \quad = \quad c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + \ldots.$$
>
> Here x is variable, each c_i is a constant. We are assuming that the power series actually represents a number dependent on x. From the way we take derivatives of series, term by term, we have
>
> $$f'(x) \quad = \quad 0 + c_1(1) + c_2(2)x + c_3(3)x^2 + c_4(4)x^3 + \ldots.$$
>
> This derivative has a derivative. The derivative of the derivative is called the **second derivative**. It is denoted by $f''(x)$, given by
>
> $$f''(x) \quad = \quad 0 + 0 + c_2 2(1) + c_3 3(2)x + c_4 4(3)x^2 + \ldots.$$
>
> Clearly, that is not the end of it. The derivative of the second derivative is the **third derivative**
>
> $$f'''(x) \quad = \quad 0 + 0 + 0 + c_3 3(2)1 + c_4 4(3)2x + \ldots,$$
>
> and the recursion continues.

It was in this context the Lagrange invented the notation f', f'', and so on. In general, we use $f^{(n)}$ to avoid $f^{\text{bunch of primes}}$. He named those things **derived functions**, which is the origin of our "derivative."

> Observe that the (derivative) series give
>
> $$f(0) = c_0, \qquad f'(0) = c_1, \qquad f''(0) = 2(1)c_2, \qquad f'''(0) = 3(2)1c_3,$$
>
> and in general
>
> $$f^{(n)}(0) \ = \ n!c_n.$$
>
> Write those the other way around:
>
> $$c_0 = f(0), \qquad c_1 = f'(0)/1!, \quad c_2 = f''(0)/2!, \qquad c_3 = f'''(0)/3!,$$
>
> and so on. Then
>
> $$f(x) \ = \ [f(0)] + [f'(0)/1!]\,x + [f''(0)/2!]\,x^2 + [f'''(0)/3!]\,x^3 + \ldots.$$

The last expression is called the **Taylor series** for $f(x)$. The import of what we have argued so far is that if a function is given by some series, then that series has to be the Taylor series. (See Boyer about Brook Taylor and why the name "Maclaurin series" is a misnomer.)

The power series we know are for $1/(1 - x)$, e^x, and sin x. Let us see the Taylor series for the sine; work on the other two in Exercises 2 and 3.

> Write $g(x) = \sin x$. We know from various exercises that
>
> $g'(x) = \cos x, \qquad g''(x) = -\sin x, \qquad g'''(x) = -\cos x, \qquad g^{(4)}(x) = \sin x,$
>
> which means the subsequent derivatives recycle through these. Therefore
>
> $g(0) = 0, \quad g'(0) = 1, \quad g''(0) = 0, \quad g'''(0) = -1, \quad g^{(4)}(0) = 0, \ldots.$
>
> The Taylor series for $g(x)$ is
>
> $g(x) = 0 + [1/1!]\,x + [0/2!]\,x^2 + [-1/3!]\,x^3 + [0/4!]\,x^4 + [1/5!]\,x^5 + \ldots.$
>
> That matches our familiar
>
> $\sin x = x - x^3/3! + x^5/5! - \ldots.$

Lagrange stood the Taylor argument on its head. Taylor had characterized the series in terms of the derivatives; Lagrange drew the derivatives from the series.

> In Lagrange's notation, write
>
> $f(h) \quad = \quad c_0 + c_1 h + c_2 h^2 + c_3 h^3 + c_4 h^4 + \ldots.$
>
> He understood that equation to give f in the vicinity of $x = 0$. More generally,
>
> $f(x + h) \quad = \quad a_0 + a_1 h + a_2 h^2 + a_3 h^3 + a_4 h^4 + \ldots,$
>
> where now the coefficients $a_i = a_i(x)$ are dependent on x. Then he *defined* the derived functions by
>
> $f'(x) = 1!\, a_1(x), \qquad f''(x) = 2!\, a_2(x), \qquad f'''(x) = 3!\, a_3(x), \ldots.$

Exercises VIII.B.4a

1. What point of the unit circle is closest to (3, 4)? (This is a **constrained-optimization problem**. It asks: Of the points that satisfy the requirement ("constraint")

   $g(x, y) = x^2 + y^2 = 1,$

   which one gives the smallest value of ("optimizes")

   $f(x, y) = \sqrt{([x - 3]^2 + [y - 4]^2)}$?)

   [Essential first step: Optimize f^2 instead. Then if you know Lagrange's method, give it a workout. Instead, you can answer using calculus. On the third hand, you could use trigonometry. Whichever you choose, check your answer against the easy geometric answer.]

2. Let $f(x) = 1/(1 - x)$. We know that as long as $-1 < x < 1$,

   $f(x) = 1 + x + x^2 + x^3 + \ldots.$

   a) Use the series to evaluate $f(0)$, $f'(0)$, $f''(0)$, ….

   b) Use Fermat's method to prove by induction that for any n,

   $f^{(n)}(x) = n!/(1 - x)^{n+1}.$

   c) Do the values $f^{(n)}(0)$ from (b) match the answers from (a)?

3. Let $g(x) = e^x$.

   a) Write the formulas for $g'(x)$, $g''(x)$, …, and the values of $g(0)$, $g'(0)$, $g''(0)$, ….

   b) Write the Taylor series for e^x.

b) the remainder

The credibility of section (a) above depends on how much faith you have that functions are given by series and that you can treat series like finite sums, as in doing derivatives term by term. Lagrange did not depend on such faith. He showed that every appropriate function [one whose derivatives are doable] is given by an ordinary sum—a *finite* part of its Taylor series—to within a describable error.

Proposition. (The Lagrange Remainder) The difference between $f(x)$ and the (finite) sum

$$[f(0)] + [f'(0)/1!]\,x + [f''(0)/2!]\,x^2 + \ldots + [f^{(n)}(0)/n!]\,x^n$$

is exactly the **remainder**

$$R_n = [f^{(n+1)}(t)/(n+1)!]\,x^{n+1}$$

for *some* (*unspecified*) value t between 0 and x.

Stay with the example of the sine function.

In the series for

$$g(x) = \sin x,$$

make $x = \pi/6$. Lagrange's proposition says that

$$\sin(\pi/6) = \pi/6 - (\pi/6)^3/3! + (\pi/6)^5/5! + R_5,$$

with the understanding that for some angle t (as yet unknown, but not after Exercise 1) between 0 and $\pi/6$, the remainder R_5 is exactly

$$[g^{(6)}(t)/6!]\,x^6 = [-\sin t]/6!\,(\pi/6)^6.$$

We can calculate

$$\pi/6 - (\pi/6)^3/3! + (\pi/6)^5/5! \approx 0.500\,002\,132\,6.$$

That result comes from a scientific calculator that displays ten significant figures [but, I think, actually calculates with thirteen]. Therefore it is likely to be accurate to at least eight decimals. Remember, though, that we can only trust that calculation to within R_5. Even without knowing the value of $[-\sin t]$—merely knowing that it is between -1 and 0—we can be sure that

$$0 > R_5 > [-1/6!]\,(\pi/6)^6 \approx -0.000\,03.$$

We can make the same calculation with

$$x = (10 + 1/6)\pi = 61\pi/6.$$

That is,

$$\sin(61\pi/6) \approx (61\pi/6) - (61\pi/6)^3/3! + (61\pi/6)^5/5!,$$

with an error of exactly

$$r_5 = [-\sin u]/6!\,(61\pi/6)^6.$$

At worst, the absolute value of that (not necessarily negative) error might be

$$[1/6!]\,(61\pi/6)^6 \approx 1.5 \text{ million}.$$

That is a most impressive possible error, given that we know $\sin(10 + 1/6)\pi = 1/2$. Indeed, the actual error is about 272000.

However, if you increase the number of terms *far enough*, then the error begins to disappear. Thus,

$$\begin{aligned} R_{2047} &= [\sin v/2048!]\,(10\pi + \pi/6)^{2048} \\ &< 32^{2048}/2048! && \text{(in absolute value)} \\ &= 1024^{1024}/[1(2)3\ldots1024(1025)1026\ldots2048] \\ &< 1/1024!. && \text{(Explain all.)} \end{aligned}$$

The subsequent error estimates are smaller still.

Exercises VIII.B.4b

1. Use the calculation

$$\pi/6 - (\pi/6)^3/3! + (\pi/6)^5/5! \approx 0.500002$$

to estimate the angle t for which

$$\sin \pi/6 = \pi/6 - (\pi/6)^3/3! + (\pi/6)^5/5! - \sin t/6!\,(\pi/6)^6.$$

2. a) In section VII.B.4b, we said that

   $e^1 \approx 1 + 1/1! + 1/2! + 1/3! + 1/4!$,

   with an error of less than 0.01. Write an expression for the exact error, meaning the Lagrange remainder R_4, then give an estimate for its value.

   d) If we calculated

   $e^1 \approx 1 + 1/1! + 1/2! + \ldots + 1/10!$,

   how big could the error R_{10} be?

---

c) limits

Now we can use d'Alembert's language to specify what it really means to write

$$\sin x = x - x^3/3! + x^5/5! - \ldots.$$

The meaning is that sin *x* is *the limit* of the sum

$$x - x^3/3! + x^5/5! - \ldots \pm x^{(\text{odd } n)}/(\textit{odd } n)! \qquad \text{(whichever sign is right)}$$

as *n* tends to infinity. In words, we can make the sum "approach" to sin *x* "nearer than by any given quantity" by forcing *n* to be correspondingly big.

In the example toward the end of section (b) above, we would find

$$[\text{sum}] - \text{sine} = [(61\pi/6) - (61\pi/6)^3/3! + (61\pi/6)^5/5! - \ldots \pm (61\pi/6)^n/n!] - \sin(61\pi/6)$$

smaller in absolute value than 1/1024! for all $n \geq 2047$.

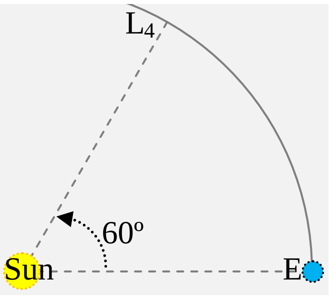
[Oddly, Lagrange *specifically excluded* any talk of limits. Maybe he thought they did not fit into an algebraic treatment. The separation has a point: In our training, the notion of limit is exactly where calculus takes over from algebra.]

d) the three-body problem

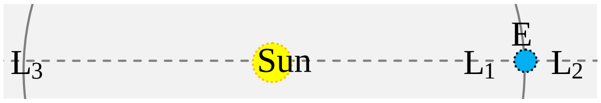
Newton's solution for the orbits of the planets—his description (section VII.B.4c(ii)) of the possible paths—depended on the assumption that the Sun sits stationary at the origin of coordinates. We admitted back there that in fact the Sun necessarily moves. That is always the case for two isolated bodies. Newton himself showed that they must both describe elliptical orbits about their center of mass (unless low speeds make them collide at the center. You can think of "center of mass" as the average position of their masses.) That center *is* stationary. At one extreme is the situation in which their masses are equal. Then the center of mass is at the midpoint of their segment, and the mutual orbiting is obvious. At the opposite extreme, one mass is much bigger than the other. Then the center is near—or even inside—the massive one, and its orbiting is almost undetectable. Thus, Earth's mass is more than 80 times the Moon's. Their center of mass is therefore about 1/80 as far from Earth's center as from Luna's. That puts the center about 3,000 miles from the center of Earth's 4,000-mile-radius sphere. The effect is even more pronounced for Sun and Earth; the mass ratio there exceeds 300,000.

Incredibly, if you add a third body, then the problem of solving the equations of motion becomes intractable. There simply does not exist a general solution that describes all the possible paths the three objects might follow in orbiting the center of mass. There are only special-case solutions. Lagrange studied some of the stable ones.

One stable arrangement is when one mass is great and the other two revolve around it independently, with small interaction. That describes the relationship of Sun, Venus, and Earth. A second stable configuration is that of Sun, Earth, and Moon: massive central object, smaller object, and even smaller third object looping around the second as they both revolve around the massive one. Lagrange's study of this latter configuration was crucial to our understanding of Luna's motion. That included the hoped-for use of lunar positions to determine longitude on Earth. The revolutionary government chose him to create the Office of (*Bureau des*) Longitudes.



A very different stable arrangement is in the figure at right. Start with Sun and Earth: massive body, smaller body. If you put a very-low-mass object at the **Lagrangian point** L_4, on (roughly) Earth's orbit 60° ahead of Earth and having Earth's orbital speed, then the object will *stay* 60° ahead of Earth. It is obvious why it should orbit the Sun at the same rate as Earth does; it has the speed needed on that orbit. The unintuitive part is that the object does not get drawn toward Earth. The combination of the three gravitational attractions makes it revolve around Earth in the same one-year period in which it orbits the Sun, all three bodies dancing about their center of mass. That keeps the object at relatively stable distance from Earth. From considerations of symmetry, it is clear that the point L_5 travelling 60° *behind* Earth is another stable position.



There are three other Lagrangian points, all on the line joining Sun and Earth, as shown at left. Those three were originally described by Euler. The point L_1 lies about a million miles from Earth, toward the Sun. Under the influence of just the Sun, an object at that place would orbit Sun with a period smaller than a year. (By Kepler's Third Law, its period would be about $[92/93]^{3/2}$ of Earth's.) With Earth attracting to the outside, the object needs to orbit less fast; its period increases to Earth's period. Therefore an object there remains on the Sun-Earth segment, roughly constant distance from both. At the two points L_2 and L_3, respectively about a million miles outward from Earth and just beyond the point opposite Earth, the attractions of Sun and Earth reinforce. The sum of attractions forces an object at those points to orbit faster; the sum *decreases* the object's period to Earth's period. That keeps the object stable behind or opposite Earth.

Mankind has put numerous satellites at Sun-Earth's L_1 and L_2 points. (See NASA for discussion of the nature of the points—including for example their actual stability—and some of the satellites humans have put there.) There has been no great reason for putting anything at L_4 and L_5, and it would be nearly impossible to communicate with something at L_3. But L_1 and L_2 offer advantages. From L_1, a satellite has a perpetual, unobstructed view of the Sun. [That's discounting the occasional transit of Mercury or Venus, for which events those satellites enjoy an enviable seat.] Turning around, such a satellite has a constant view of Earth's whole sunlit side, interrupted now and again by the Moon. At L_2, a satellite would see Earth's dark half, but in the outer direction would have a shaded look at the starry background. [The point is not in perpetual darkness. From 1 million miles, Earth's angular size is about $8{,}000/1\text{M} = 0.008$ radian $\approx 0.46°$. From the corresponding 94M miles, Sun's angular size is approximately $865\text{K}/94\text{M} \approx 0.009$ radian. From L_2, you can always see some of the Sun.]

5. Gauss and Probability

Carl Friedrich Gauss [rhymes with "house"] (1777-1855) and Euler are generally recognized as the most important mathematicians since the giants of Greek fame. Much of nineteenth century mathematics came from Gauss, or flowed out of his work, or was anticipated by ideas he chose not to publish. We can sneak him into the eighteenth century because of epochal results he produced by 1801.

["Giant[s] of Greek fame" is not my invention. Read Emma Lazarus's poem.]

Gauss was a legend practically from childhood. There is a tale that as a child, he amazed his teacher by instantly summing the integers from 1 to 100. The teacher had given that task to the class to fill some time. Presumably, The Kid realized that the sum consists of 1 + 100, 2 + 99, …, 50 + 51, a parade of 50 pairs adding to 101 each, and he was good at multiplication as well. At the age of 18, he proved that it is possible to construct a regular 17-gon, and more generally polygons whose sides number certain combinations of Fermat primes. (Those are of the form $2^{2^n} + 1$; see Section VII.A.4f(iii).) At 22, he presented a doctoral thesis with the first proof of one of algebra's most important theorems (which we will cover). In later life, his analytical description of surfaces established a whole new branch of geometry. His discoveries in PDE's, and in connecting them to phenomena of electricity and magnetism, began the development that led to James Maxwell's complete description of electromagnetism.

From his work in analysis, the part we can most easily describe actually belongs to probability. It is the normal approximation to the binomial distribution.

a) the binomial distribution

Imagine a set of actions that terminates with either of two possible results, with predictable probabilities. For example, one roll of the dice in the casino game of "craps" will produce a WIN or a LOSS for you, with (calculable) probabilities roughly 0.493 and 0.507 respectively. In a single roll, then, you have a probability .493 of ending up with a profit.

[In the study of probability, the two possible results are always called "success" and "failure." We will stick with the roller's point of view and WIN-LOSS.

A "roll" in craps is actually a sequence of throws of a pair of dice. Known rules specify when the sequence ends and whether you win. It is not hard to classify the infinite number of possible sequences and thereby arrive at the 0.493 probability of WINning. See the description at MathForum.org.

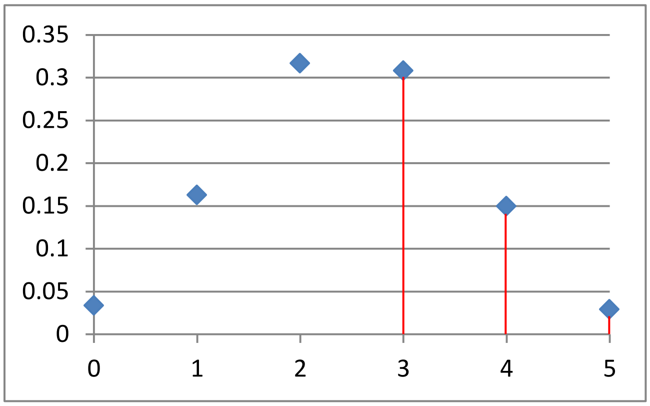
It *is* embarrassing to write a sort of history of math and leave untouched the history of probability. This particular detour into probability is worth taking.]

Let us perform "sessions" with stated numbers of rolls and count how many WINs happen.

> Suppose you decide to do a session of five rolls. Assuming the playing components—the dice, the playing table, the drinks—are not defective, your probability of WINning any given roll is **independent** of what happened before: It *stays* 0.493, irrespective of WINs or LOSSes on any previous rolls. The rolls constitute **independent trials**. (The name **Bernoulli trials** is synonymous, as long as we specify in advance how many rolls there will be. Jacques was first to describe them.) Given the independence, the probability of your WINning 0, 1, 2, 3, 4, or 5 of the rolls is given, respectively, by the six terms of the binomial expansion
>
> $$(.507 + .493)^5 \;=\; .507^5(.493^0) + \binom{5}{1}.507^4(.493^1) + \ldots + \binom{5}{4}.507^1(.493^4) + .507^0(.493^5).$$
>
> [Online, plenty of pages name the formula for those terms, but I could not find an online explanation why they give the probabilities. Here is a quick one. The probability of losing the *first* roll, then winning the next four, is
>
> $$.507 \times .493 \times .493 \times .493 \times .493.$$
>
> But that is not the only way to win exactly four rolls. You can pick any four of the rolls, win those four, lose the other one. Because there are $\binom{5}{4}$ distinct ways to pick a "combination" of four objects from a set of five, there are that many ways to name four rolls to win. Therefore the probability of winning four rolls out of five is $\binom{5}{4}.507^1.493^4$.]

The values of the terms are given in the table below and plotted on the chart at right. (Verify any of them.) From the connection to the binomial expansion, the function given by the table and graphed in the chart is called the **binomial distribution**. Observe that for the five-roll session, your probability of turning a profit—meaning getting 3, 4, or 5 WINs—is reduced to

$$.308 + .1497 + .0291 \approx .487.$$



0	1	2	3	4	5
0.0335	0.1629	0.3168	0.308	0.1497	0.0291

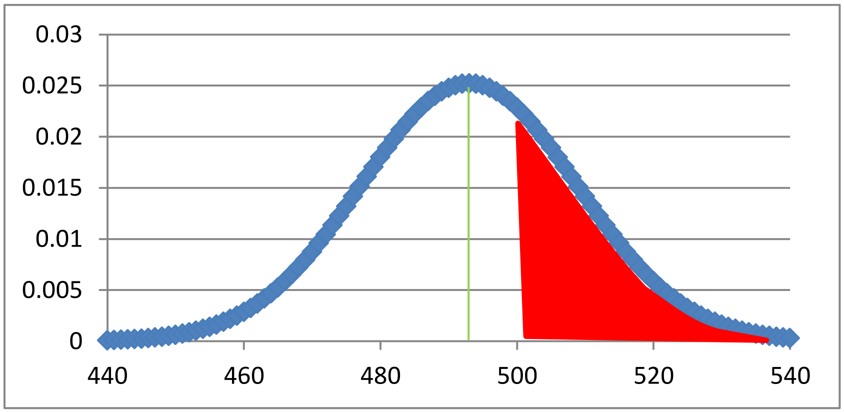
It will be handy to have some technical terms. The sequence LWWWW is an **outcome**. We saw above that its probability is $.507^1(.493^4)$. The set

$$E = \{\text{LWWWW, WLWWW, WWLWW, WWWLW, WWWWL}\}$$

of outcomes is an **event**, with probability $\binom{5}{4}.507^1(.493^4)$. It is a special event, in that all of its outcomes have the same number of WINs. Our interest is a function w that assigns a number to each such event, namely the shared number of WINs in the event's outcomes. Thus, $w(E) = 4$. Such a function is a **random variable**. Its values inherit probabilities. We say, for example, that $w = 4$ has the probability we just named, $\binom{5}{4}.507^1(.493^4)$. Because the possible values of w are separate, w is a **discrete** random variable. (The number of possible values need not be finite; see Exercise 1.)

On the plot, we see the three red verticals erected at $x = 3$, 4, and 5. The sum of their lengths also gives the probability of a majority WINs. In the example of five rolls, there is no reason to add those lengths. But there is good reason to do it when the number of rolls is large.

> Indulge yourself: Stretch the session to 1000 rolls. [That's not superhuman. It is easy to make 100 bets per hour at the craps table. A weekend in Las Vegas will easily yield 1000 rolls.]
>
> The corresponding probability data are plotted at right, truncated to the interval from 440 to 540 WINs. (Do Exercise 2b-d to see that below 440 and above 540 are not worth showing.) For the complete chart of 1001 data points, the heights are given by the terms of the expansion
>
> $$(.507 + .493)^{1000} = .507^{1000} + (1000)\,.507^{999}(.493) + 1000(999)/2!\;.507^{998}(.493^2) + \ldots + .493^{1000}.$$
>
> The highest point is at 493 WINs (Exercise 2a), where the green vertical is; we will come back to that value later.



Your probability of emerging from the 1000-roll session with a profit is the sum of the red verticals from 501 to 1000. That collection of verticals is hard to tell from an area, as indeed the graph is hard to distinguish from a continuous one.

> Pretend that it *is* continuous, then think like Wallis. That is, think of the red verticals as the infinitesimals for the part of the area under the graph from $w = 501$ to $w = 1000$. Wallis would tell us that the quotient
>
> (sum of red verticals)/(sum of all 1001 verticals)
> $= $ (probability of profit)/$(.507 + .493)^{1000}$
> $= $ probability of profit
>
> is the fraction of the area under the whole graph that lies between $w = 501$ and $w = 1000$.

Notice that we discussed the probability of profit, but did not evaluate it. Evaluation comes later.

We will exploit the connection to areas. We are studying a random variable whose possible values are separate. We will pursue that study by investigating variables whose possible values form a continuum of real numbers.

Exercises VIII.B.5a

1. The **geometric distribution** applies to "trials until success." Think of a die whose probability of landing with SIX showing is (always) 1/6. Imagine rolling it repeatedly until SIX shows; that is the "until success" part. Thus, the outcomes can be 1, 2, 3, … (rolls). Let $p(n)$ represent the probability of (SIX showing *for the first time* on roll number) n.
   a) Evaluate $p(1)$, $p(2)$, $p(3)$, …. (These rolls are also independent trials. Therefore
   [probability of this followed by that] = [probability of this] × [probability of that].)
   b) Add up that infinity of probabilities.
   c) Find by two methods, adding and multiplying, the probability that 3 or more rolls will be needed for a SIX to show.
2. Let $p(n)$ represent the probability of n WINs in 1000 rolls of the dice, each roll having probability 0.493 of a WIN. Show that:
   a) $p(0) < p(1) < \ldots < p(493)$ and $p(493) > p(494) > \ldots > p(1000)$.
   (Hint: Write out the binomial coefficients, then simplify the ratios
   $p(0)/p(1), \ldots, p(492)/p(493), p(493)/p(494), \ldots, p(999)/p(1000)$.
   No calculation is necessary.)
   b) $p(449)/p(450) < 0.84$.
   c) $p(434)/p(435) < p(435)/p(436) < \ldots < p(449)/p(450)$.
   d) $p(434) < 0.062\,p(450)$. Any calculator can do the needed power. This relation suggests why the probabilities below 440 WINs are too small to plot; similarly above 540.

b) continuous distributions

We have seen probabilistic results that are integers, possibly an infinite number of them. There are, however, *common* probabilistic phenomena whose results can fill an interval of the real line. An everyday such phenomenon is **waiting time**.

Imagine that you arrive at no special time at a bus stop where the buses arrive ten minutes apart on a consistent schedule. The time you have to wait for the next arrival is therefore some real number between 0 and 10 minutes. In this case, the function that assigns to each waiting session its duration has possible values that span a real interval. Such a function is a **continuous random variable**.

Of necessity, the probability of hitting any exact waiting time is zero. It makes sense to talk, instead, about a value's falling into some subinterval of the interval of possible times. (A subinterval is a set of possible outcomes; it is an *event*, with some probability.) As we hinted above, with "Wallis infinitesimals," we turn to areas under graphs. We will specify a function f called the **probability density function**.

> It is density in the sense that around $t = a$, $f(a)$ is *probability per unit of length* on the t-axis. That is, for the infinitesimal length between
>
> $t = a$ and $t = a + dt$,
>
> the probability of landing there is
>
> $dp = f(a)dt$.
>
> Then the probability that a value falls into the interval from $t = a$ to $t = b$ is the summation of the infinitesimals dp. In other words, it is the integral of $f(t)$ between those values of t. In our language, it is the area under the corresponding part of the graph of f.

There is one restriction we place on the density. Recall that the verticals under our two previous plots added up to $(.507 + .493)^5$ and $(.507 + .493)^{1000}$, both 1. Accordingly, we will insist that the area under the complete graph of $f(t)$ must be exactly 1.

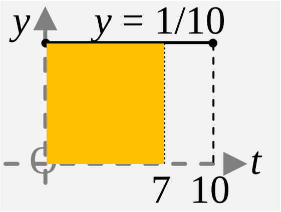
(i) the uniform distribution

View the example of arriving to catch the buses that come every ten minutes. Our intuitive idea of randomness suggests that you are as likely to arrive during the third minute after the previous bus, $t = 2$ to $t = 3$, as during the seventh, $t = 6$ to $t = 7$. In other words, you are as likely to wait 7-8 minutes as to wait 3-4 minutes. [Experience, of course, tells us that we will barely miss a bus and have to wait 9.9 minutes.] More generally, any two wait-intervals of equal length have equal probabilities. That situation is governed by the uniform distribution.

For the **uniform distribution**, the density function is always (a correctly scaled) constant. The corresponding graph is given by

$$y = f(t) = \text{constant } c \qquad \text{for } r \le t \le s.$$

For the bus example, $r = 0$ and $s = 10$. To make the total area $10c = 1$, we have to take $c = 1/10$, shown at right. Then, for example, the probability that your waiting time falls between 0 and 7 minutes is the orange area, 0.7. You are unlikely to wait more than 7 minutes.



(ii) the exponential distribution

A different example of waiting involves radioactivity. In a radioactive element, atoms break down at random intervals, emitting some of their constituent particles. (Compare Exercise VII.B.4c:3.) If you surround a sample of say radium with detectors—like Geiger counters or phosphorescent screens—then you can detect the emissions. Physical considerations (later) suggest that from when you start looking, the time you wait for the next emission has the **exponential distribution,** a scaled version of the graph

$$y = g(t) = e^{-t}, \qquad t \ge 0.$$

Go wait for the bus. Imagine that the dispatcher is sending buses, still at an *average* rate of six buses every sixty minutes, but not at precise ten-minute intervals. Instead, he dispatches a bus whenever his radium sample emits a particle. Accordingly, the buses are separated by exponentially-distributed gaps.

> [For some perspective, recall from the radium exercise that the mass m of a radium sample decreases at the rate *per year* of
>
> $dm/dt = -0.00753\, m$.
>
> A radium atom has a mass of about 3.8×10^{-17} gm. For the sample to lose six atoms (six times that mass) in
>
> 1 hour $= 1/(24 \times 365.24)$ year ≈ 0.000114 yr,
>
> it has to start with
>
> $$m = dm/(-0.00753\, dt) = -6 \times 3.8 \times 10^{-17} \text{ gm}/(-0.00753 \times .000114)$$
> $$\approx 2.7 \times 10^{-10} \text{ gm.]}$$

From mathematical considerations (also later), it follows that the scaling

$$y/d \;=\; e^{-t/c}, \qquad t \ge 0 \text{ minutes},$$

has to have c = 10. The area under the graph of

$$y = e^{-t/10} \qquad \text{from } t = a \text{ to } t = b$$

is $10(e^{-a/10} - e^{-b/10})$ (Exercise 2). Setting a = 0 and b = ∞, which Wallis would approve, we find the total area under the graph to be

$$10(e^{-0} - e^{-\infty}) \;=\; 10.$$

Therefore the scaling factor d has to be 1/10.

We see that the density is

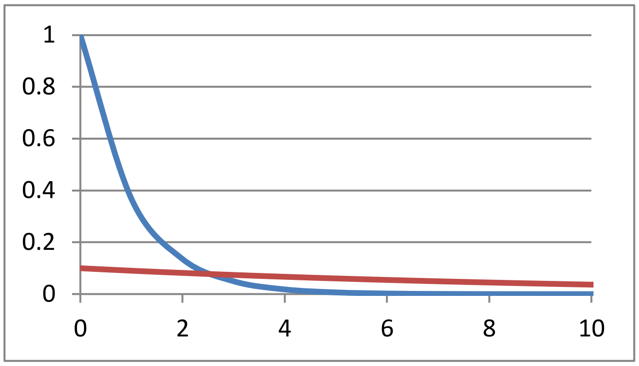
$$g_{10}(t) \;=\; (1/10)e^{-t/10}, \qquad t \ge 0.$$

The probability of waiting between a and b minutes for one of the random buses is the area under the graph of g_{10} from $t = a$ to $t = b$, an area equal to $(e^{-a/10} - e^{-b/10})$.

How likely is it that your wait will be 0-7 minutes? The probability is

$$e^{-0} - e^{-0.7} \;\approx\; 0.503.$$

You are about even money to wait more than seven minutes.



It is worthwhile to compare the graphs of

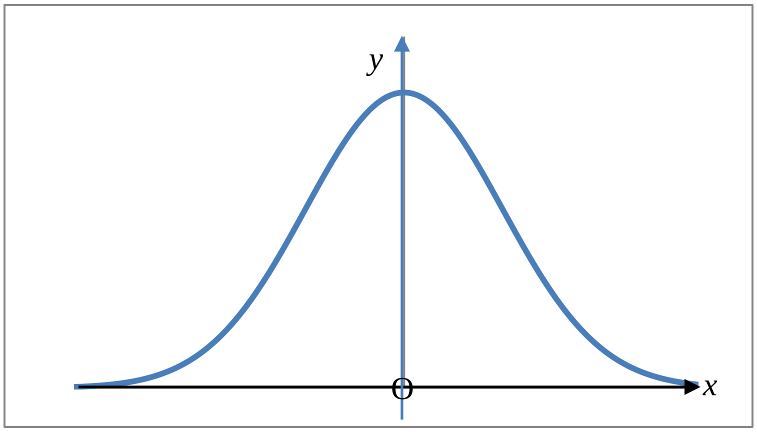
$$y = e^{-t} \qquad \text{and} \qquad y = (1/10)e^{-t/10}.$$

The two are sketched at right, blue and red respectively; see Exercise 1 for justification. Notice that the latter starts lower and goes down less fast. A greater part of the area under it is away from x = 0.

One of the physical considerations we mentioned is that the radium atoms cannot remember when the last one popped. In other words, if the last emission was 15 minutes ago, then the probabilities for further *additional* wait-times are the same as for wait-times from the last emission. In that situation, we say the intervals between emissions **are independent**. In any situation where you can make that statement, the exponential distribution applies.

To see independent wait-times in familiar terms, suppose you have been waiting 15 minutes for a bus when another rider arrives at the stop. The radium atoms, unaware that he is a newcomer, offer him the same distribution of wait-times they offered you. Thus, he has probability 0.503 of waiting 0 to 7 minutes, 0.497 of waiting more. Since you will stand there as long as he does, the probability that you have to wait *another* 7 minutes or more is unsympathetic to how long you have already stewed.

(iii) the normal distribution



Now go back to the figure in (a) of the binomial distribution of WINs for 1000 rolls in the game of craps. Replace the 1001-point pattern by a smooth curve, and you produce the graph at right. That bell-shaped graph, "the curve," goes with the normal distribution.

The curve has the shape of

$$y \;=\; e^{-x^2},$$

with no restriction on x. We can check

(Exercise 3) that the latter graph has its highest point at $x = 0$, horizontal tangent there, and tangents of negative slope for $x > 0$. Because

$e^{-x^2} = 1/e^{x^2} < 1/x^2$ (Justify both.),

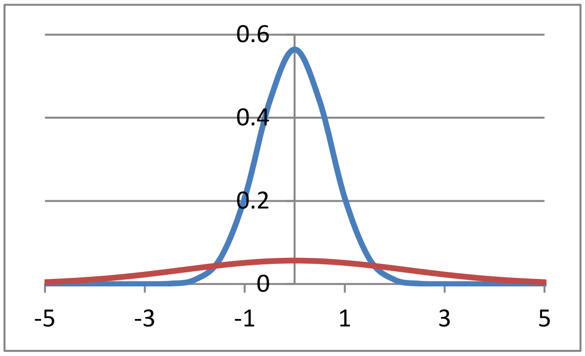
the positive x-axis is an asymptote to the graph toward the right. The left ($x < 0$) half is symmetric to the right half. The curve in the figure is the simplest we can draw to fit that information. Properly scaled, it is the density function for the **normal distribution**. Gauss proved that as the number of Bernoulli trials increases toward infinity, the binomial distribution approaches a normal distribution centered at the binomial maximum.

> The scaling takes the form
>
> $$y/d = e^{-(x/c)^2}.$$
>
> You need multivariable calculus to determine the area under that graph, $cd\sqrt{\pi}$. To make the area 1, we must set $d = 1/(c\sqrt{\pi})$. Therefore the density function for a normal distribution is
>
> $$h(x) = (1/[c\sqrt{\pi}])e^{-(x/c)^2}.$$

As with the exponential, compare the graphs of

$y = (1/\sqrt{\pi})e^{-x^2}$ and $y = (1/[10\sqrt{\pi}])e^{-(x/10)^2}$.

They are illustrated at right, again blue and red, and related as before: The second one starts lower and drops more slowly to right and left. Accordingly, a greater share of the area under it is away from the high point.



Exercises VIII.B.5b

1. a) Sketch the graph of $y = e^{-t}$ for $t \geq 0$.
   (Hint: The graph of $y = e^{-t}$ is the left-for-right mirror image of the graph of $y = e^t$.)
   b) Show that
   $e^{-t}/([1/10]e^{-t/10})$
   is 10 when $t = 0$, decreases as t increases, is 1 at some t, and is 0 when $t = \infty$.
   c) Use (a) and (b) to sketch the graph of $y = (1/10)e^{-t/10}$, $t \geq 0$.
2. Think of areas under the graph of
   $y = G(t) = e^{-t/10}$, $t \geq 0$
   in Wallis's infinitesimal terms.
   a) Draw the verticals from the t-axis up to the graph at $t = a$, $t = b$, and the midpoint
   $t = (a + b)/2$.
   Show that their heights are those of the verticals under the graph of
   $y = g(t) = e^{-t}$
   at the places $t = a/10$, $t = b/10$ and $t = (a + b)/20$.
   b) Evaluate the area under
   $y = g(t)$, $t = a/10$ to $t = b/10$.
   (Hint: See the hint for Exercise 1a, and remember that the antiderivative of e^t is e^t.)
   c) Part (a) suggests that there are equal corresponding infinitesimals under the graph of $G(t)$, $t = a$ to $t = b$, and the graph of $g(t)$, $t = a/10$ to $t = b/10$; but the ones for G are spread out over a *width* ten times as big as the infinitesimals for g. In view of (b), what does that imply for the area under G?

3. a) Use Fermat's method to show that the slope of the tangent to the graph of

$$y = e^{-x^2}$$

at the point (a , e^{-a^2}) is -2ae^{-a^2}. (Hint: If h and k are not zero, then

$$e^{hk}/h = ke^{hk}/kh,$$

and if s is infinitesimal, then

$$[e^s - 1]/s = 1.)$$

b) Use the slope to show that the graph rises to the right for $x < 0$, has horizontal tangent at $x = 0$, and drops to the right for $x > 0$.

---

c) average value and average dispersion

Any probability distribution carries numerous important parameters. The two most basic are the average value of the associated variable and the dispersion, the extent to which the values are scattered away from the average.

[To get an imprecise analogy, think of planning a trip. You would want to know the average temperature of destinations you consider. You might prefer a place with a 75° average to one with a 30° or 98° average. Still, if you choose the 75°, you should also look into the dispersion. If Honolulu averages 75° with a daytime high of 82° and overnight low 68°, and Death Valley has the same average with high 115° and low 35°, you might judge the places unequally attractive.]

(i) expected value

Let us go back to the craps table and look at a bunch of customers playing five rolls each. Based on the probabilities in the five-roll data in (a), we would calculate that among the customers, a fraction

0.0335	will have 0 WINs,
0.1629	will have 1 WIN,
…,	
0.0291	will have 5 WINs.

With those fractions, we anticipate that the *average* number of WINs per customer will be

$$(0.0335)[0] + (0.1629)[1] + (0.3168)[2] + (0.308)[3] + (0.1497)[4] + (0.0291)[5] = 2.465.$$

For any discrete variable, no matter what the distribution is, the summation of all the products

(probability of result)[value of result]

is called the **expected value** or **expectation**. If the random variable in question has an infinity of values, then the summation becomes an infinite series (and the expectation might be infinite; see Exercise 1).

[Remember that "expected value" is a strictly technical name. "Average value" would be fine with me. If I rolled the dice five times and scored 2.465 WINs, I would refer to the result as "unexpected."

If the variable has infinitely many positive values and infinitely many negative, then a mathematical complication arises. We won't even try to address it.]

Suppose we have instead a continuous variable with probability density $h(x)$. The discrete definition

expected value = summation of (value × probability)

yields naturally to

expected value = integral of (value × probability)
= area under the graph of $y = xh(x)$.

See Exercise 2 for the expected values of the continuous distributions we have described. For the uniform distribution (2a), the integral is at our level. The same is true for the normal (2c), for a reason you should see for yourself. For the exponential (2b), undergraduate-level calculus is unavoidable.

Exercises VIII.B.5c(i)

1. Recall the probabilities (Exercise VIII.B.5a:1) $p(1)$, $p(2)$, … of needing 1, 2, … throws of a fair die "until success," meaning until a SIX shows up.

   a) What is the expected value of the number of throws needed? (Hint:

   $$1 + 2a + 3a^2 + 4a^3 + \ldots \;=\; (1 + a + a^2 + a^3 + \ldots) + (a + a^2 + a^3 + a^4 + \ldots) + (a^2 + a^3 + a^4 + a^5 + \ldots) + \ldots.$$

   The result is not a coincidence. Try to show that the geometric distribution's expected value is always related that way to the probability of "success.")

   b) Suppose the King, grateful for your service, lets you roll that die with the promise that if the first SIX happens on roll number n, he will reward you with 2^n dollars. What is the expected value of your reward?

2. Find the expected values for:

   a) the uniform distribution with density

   $f(x) = 1/10$, x between 0 and 10.

   Then show that in general, for the uniform distribution between $x = a$ and $x = b$, the expected value is the average (arithmetic mean) $(a + b)/2$.

   b) (calculus) the exponential distribution with

   $g_{10}(x) \;=\; (1/10)e^{-x/10}$, $x = 0$ to $x = b$.

   Then set $b = \infty$ to find the expected value for the unlimited interval. This answer is why, if the buses average one every 10 minutes, then the denominator under x has to be 10.

   c) (only *seemingly* calculus) the normal distribution with

   $h_{10}(x) \;=\; 1/(10\sqrt{\pi})\, e^{-(x/10)^2}$, over the whole real line.

(ii) standard deviation

To describe how widely spread the values of a random variable are, we could find the average difference between the values and the expected value.

Return to the craps table. For our five rolls, the numbers 0, 1, …, 5 of WINs differ from the expected value 2.465 by

$0 - 2.465$, $1 - 2.465$, …, $5 - 2.465$.

The average difference, *weighted according to the frequencies* (probabilities) of the values, is

$$\begin{aligned}&(0.0335)[0 - 2.465] + (0.1629)[1 - 2.465] + \ldots + (0.0291)[5 - 2.465]\\ &= (0.0335)\,[0] + \ldots + (0.0291)\,[5] \quad - \quad (0.0335 + \ldots + 0.0291)[2.465]\\ &= \text{expected value} \quad - \quad (1)[2.465] \quad = 0.\end{aligned}$$

Of course it is zero! That is the nature of arithmetic averages, including weighted ones: For every unit of excess above the average, there must be a unit of deficiency below it.

The negative differences are cancelling out the positive ones. To avoid that misinformation, there are two ways to eliminate the signs. One is to use the absolute values of the differences. The other is to square them. The latter is how Gauss measured the average separation from expectation.

Take the squares of the differences,

$(0 - 2.465)^2$, $(1 - 2.465)^2$, …, $(5 - 2.465)^2$.

The average of the squares, weighted by probabilities, is

$$(0.0335)[0 - 2.465]^2 + (0.1629)[1 - 2.465]^2 + \ldots + (0.0291)[5 - 2.465]^2.$$

The square root of that number,

$$\sqrt{(0.0335)[0 - 2.465]^2 + \cdots + (0.0291)[5 - 2.465]^2} \;\approx\; 1.118$$

is the Gauss average. Notice that it looks vaguely like an extension of the distance formula.

[The square root of the average square is sometimes called the RMS, **root-mean-square**. ("RMS average" would be redundant.) Notice that the name comes from our algebraic notation, where the last operation done appears on the left. Given the order in which the operations are performed, perhaps the name ought to be "square-mean-root."

It is instructive to compare the RMS with the average distance, what you would get if you followed the first idea and averaged the absolute differences. That would be

$$(0.0335)|0 - 2.465| + (0.1629)|1 - 2.465| + \ldots + (0.0291)|5 - 2.465| \approx 0.937,$$

smaller than RMS. It is in the nature of RMS to lean toward the bigger of the numbers being averaged.]

For every random variable, the RMS distance of the values from the expected value is called the **standard deviation**. It is universal to signify expected value by μ (Greek lowercase letter *mu*) and standard deviation by σ (lower *sigma*). For a discrete variable, the deviation is

$$\sigma \quad = \quad \sqrt{}(\text{summation of the products } \{[\text{value} - \mu]^2 \times (\text{probability of value})\}).$$

That summation might be a series, whose value may or may not then be finite. For a continuous variable with density $f(x)$, the sum morphs into an integral:

$$\sigma \quad = \quad \sqrt{}(\text{integral of the products } \{[\text{value} - \mu]^2 \times (\text{probability of value})\})$$
$$= \quad \sqrt{}(\text{area under the graph of } y = [x - \mu]^2 f(x)).$$

The definition can be clumsy to apply. There is a more convenient expression for the deviation.

Theorem 1. Suppose v is a random variable with expectation $\mu(v)$. Let $\mu(v^2)$ be the expectation of v^2, the variable which assigns to each event the square of the value v assigns. If the two expectations are finite, then the standard deviation of v is

$$\sigma(v) = \sqrt{}[\mu(v^2) - \mu(v)^2].$$

In case you think those terms inside the radical are equal, calculate for the five-roll WINs w. There,

$$\mu(w^2) = (0.0335)[0]^2 + (0.1629)[1]^2 + \ldots + (0.0291)[5]^2 \approx 7.326.$$

We earlier evaluated $\mu(w) = 2.465$, whose square is about 6.1.

The new formula does accord with the earlier estimate for standard deviation:

$$\sigma(w) = \sqrt{}[\mu(w^2) - \mu(w)^2] = \sqrt{}(7.326 - 2.465^2) \approx 1.118$$

We will indicate the proof of Theorem 1 for a discrete variable. The method adapts easily to a continuous variable (Exercise 1).

Let v have the values $v_1, v_2, \ldots$ with corresponding probabilities $p_1, p_2, \ldots$. Write μ and σ for its expectation and deviation. By definition,

$$\sigma^2 = p_1 (v_1 - \mu)^2 + p_2 (v_2 - \mu)^2 + \ldots.$$

Multiply out the squares and rearrange the terms to write

$$\sigma^2 = [p_1 v_1^2 + p_2 v_2^2 + \ldots] - 2[p_1 v_1\mu + p_2 v_2\mu + \ldots] + [p_1 \mu^2 + p_2 \mu^2 + \ldots].$$

That first bracket is the expected value of v^2. The second bracket factors as

$$\mu [p_1 v_1 + p_2 v_2 + \ldots] = \mu\mu.$$

The third factors as

$$\mu^2 [p_1 + p_2 + \ldots] = \mu^2 [1].$$

Provided the expectations are real numbers, we have proved

$$\sigma^2 = \mu(v^2) - 2\mu^2 + \mu^2 = \mu(v^2) - \mu^2.$$

The above argument used silently a fact that deserves attention.

We said that

$p_1 v_1^2 + p_2 v_2^2 + \dots$

is the expected value of v^2. That expression sums terms of the form

(probability of value of v)[value of v^2].

The definition of expectation demands terms of the form

(probability of value of v^2)[value of v^2].

The two summations are equivalent. If v takes on the values 7 and -7, then

(probability of $v = 7$)[7^2] + (probability of $v = -7$)$[-7]^2$

appears in the former summation, while

(probability of $v^2 = 49$)[49]

appears in the latter. Those two are equal, because

(probability of $v^2 = 49$) = (probability of $v = 7$) + (probability of $v = -7$).

In turn, that equality is an elementary property of probability (see "Addition Rule" at Glasgow).

See Exercise 2 for the deviations of the continuous distributions of interest to us.

Exercises VIII.B.5c(ii)

1. Let v be a continuous random variable with probability density $F(x)$ over some interval of the real line. Let μ be its expected value, $E(v^2)$ the expected value of v^2, both assumed to be finite. Show that the standard deviation σ of v is given by

   $\sigma^2 = E(v^2) - \mu^2.$

   (You may take it for granted that the integral of a sum is the sum of the integrals.)

2. Use Exercise 1 and the expectations from Exercise VIII.B.5c(i):2 to find the standard deviations for:

   a) the uniform distribution with density

   $f(x) = 1/10,$ x between 0 and 10.

   Then show that in general, for the uniform distribution between $x = a$ and $x = b$, the standard deviation is $(b - a)/\sqrt{12}$. (Notice that it exceeds $[b - a]/4$. This latter number, 1/4 the length of the interval, is average distance from the midpoint of the interval.)

   b) (calculus) the exponential distribution with

   $g_{10}(x) = (1/10)e^{-x/10},$ $x = 0$ to infinity.

   (Notice that the larger the denominator, the larger the deviation. Larger denominator brings more spread; deviation measures spread.)

   c) (calculus, but not multivariable) the normal distribution with

   $h_{10}(x) = 1/(10\sqrt{\pi})\, e^{-(x/10)^2}$ over the whole real line.

   (You should get $\sigma = 10/\sqrt{2}$. Again, bigger denominator goes with greater dispersion.)

(iii) the binomial parameters

For the binomial distribution, there are simple formulas for both expectation and deviation.

Theorem 2. For the binomial distribution applying to n trials with success probability p,

$\mu = np$ and $\sigma^2 = np[1 - p].$

Use the five-roll session at craps as evidence.

We need to write out the binomial coefficients. To make room, abbreviate the WIN probability 0.493 by p, the LOSS probability 0.507 by q. Then

$$\begin{aligned}\mu(w) &= q^5\,[0] + 5/1\; q^4p\,[1] + 5(4)/1(2)\; q^3p^2\,[2] + 5(4)3/1(2)3\; q^2p^3\,[3] + \\ &\qquad 5(4)3(2)/1(2)3(4)\; qp^4\,[4] + 5(4)3(2)1/1(2)3(4)5\; p^5\,[5] \\ &= 5p\,[q^4 + (4)/1\; q^3p + (4)3/1(2)\; q^2p^2 + (4)3(2)/1(2)3\; qp^3 + (4)3(2)1/1(2)3(4)\; p^4] \\ &= 5p\,[q+p]^4 \quad = \quad 5p.\end{aligned}$$

The expected value for n rolls is np.

For the standard deviation, we work with Theorem 1. The expected value of w^2 is

$$\begin{aligned}\mu(w^2) &= q^5\,[0^2] + 5/1\; q^4p\,[1^2] + 5(4)/1(2)\; q^3p^2\,[2^2] + 5(4)3/1(2)3\; q^2p^3\,[3^2] + \\ &\qquad 5(4)3(2)/1(2)3(4)\; qp^4\,[4^2] + 5(4)3(2)1/1(2)3(4)5\; p^5\,[5^2] \\ &= 5p\,\{q^4\,[1] + (4)/1\; q^3p\,[2] + \ldots + (4)3(2)1/1(2)3(4)\; p^4\,[5]\}.\end{aligned}$$

Separate the factor in {braces} into two sums by rewriting the values [1], [2], …, [5] as

$$[0] + [1], \qquad [1] + [1], \qquad [2] + [1], \qquad [3] + [1], \qquad [4] + [1].$$

From the green numbers, we have the first sum

$$q^4\,[0] + (4)/1\; q^3p\,[1] + (4)3/1(2)\; q^2p^2\,[2] + (4)3(2)/1(2)3\; qp^3\,[3] + (4)3(2)1/1(2)3(4)\; p^4\,[4].$$

From the blues, we get the second sum

$$q^4\,[1] + (4)/1\; q^3p\,[1] + (4)3/1(2)\; q^2p^2\,[1] + (4)3(2)/1(2)3\; qp^3\,[1] + (4)3(2)1/1(2)3(4)\; p^4\,[1].$$

Instead of calculating the two sums, look at their forms. The first sum is the expected number of WINs in *four* rolls. We have already shown that it equals $4p$. The second sum is the binomial expansion for $(q + p)^4 = 1$. Substituting, we get

$$\mu(w^2) = 5p\,\{4p + 1\}.$$

The expected value of w^2 over n rolls is $np([n-1]p + 1)$.

We have

$$\mu(w^2) = np([n-1]p + 1) \qquad \text{and} \qquad \mu(w) = np.$$

Therefore

$$\begin{aligned}\sigma(w)^2 &= np([n-1]p + 1) - (np)^2 \\ &= np(1-p).\end{aligned}$$

One other property of the binomial distribution is worth mentioning. The most likely result (**mode**, in probability-talk) is the integer closest to the expected value. In the example from (a), for 1000 rolls, the most probable number of WINS (with probability barely above 0.025) proved to be (.493)1000.

(iv) the normal parameters

In all our discussion, the normal densities have been centered at $x = 0$. In use, the usual situation is a dataset with normal distribution centered at some "mean" μ. Translating the center of the distribution to $x = \mu$ is just a matter of writing the density as

$$h(x) = 1/[c\surd\pi]\; e^{-([x-\mu]/c)^2}.$$

It is straightforward (calculus, but not really) to show that this density has expected value μ. Without calculus, we can see geometrically that the dispersion remains as before. Adapting Exercise 2c above, we see that the standard deviation is $\sigma = c/\surd 2$. Therefore the all-purpose normal density is, in terms of its expectation μ and deviation σ,

$$H(x) \;=\; 1/[\sigma\sqrt{2\pi}]\; e^{-([x-\mu]/[\sigma\surd 2])^2}.$$

Entertain this question: What is the probability that a US male picked **at random** (every male equally likely to be chosen) will have height between 6′ and 6′2″?

> Suppose we know that the average height is 70″. All we may conclude is that such men are taller than average. We need to know the variation. The standard deviation would help; assume that it is 3″. It turns out we still need to know the distribution. (Compare Exercise 1.)
>
> Assume that height of US males is normally distributed. In view of the mean and deviation, the density is
>
> $$H_1(x) \;=\; (1/[3\sqrt{2\pi}])e^{-([x-70]/[3\sqrt{2}])^2}.$$
>
> To answer the question, we need the area under its graph from $x = 72$ to $x = 74$.

Rather than taking those complicated densities head-on, make the task easier by thinking in terms of *displacement from the mean*, measured *in standard deviations*. The quantity $(x - \mu)$ is (signed) distance from the expectation. Then

$$z = (x - \mu)/\sigma$$

is signed distance from μ, measured in standard-deviation units. With any normal density, the probability that a value lies between $z = a$ and $z = b$ *deviations* from the expected is the area, between those two values, under the graph of

$$H_0(z) = (1/\sqrt{2\pi})e^{-(z/\sqrt{2})^2}.$$

The density H_0 is called the **standard normal density**. It has expectation 0 and deviation 1. (For those, adapt Exercises (i):2c and (ii):2c.) Notice that it is a stationary target; the expression has no parameters. Accordingly, people make tables of areas under it, just as they make trigonometric tables. (See an unusually nice table at mathisfun.com, plus try out the interactive bell curve there. Be sure to read the instructions above the table to see what it tabulates.)

> The question above asked for the probability of a height between 72″ and 74″. Those heights are 2 to 4 *inches* above mean. Interpret that as between
>
> $$z \;=\; 2/3 \;\approx\; 0.67 \qquad\text{and}\qquad z \;=\; 4/3 \;\approx\; 1.33$$
>
> *deviations* above expectation. From the table entries, the requisite area is
>
> $$0.4082 - 0.2486 \;=\; 0.1596.$$
>
> Just under 16% of our males are that tall.

Exercises VIII.B.5c(iv)

1. Imagine we have evidence that the average New York City high temperature for July 1 is 87°F. Consider this question: How unusual is a July 1 on which the high temperature in NYC exceeds 96°? Can you answer it:
   a) with no further information?
   b) given that the highs are distributed uniformly between 75° and 99°?
   c) given instead that the standard deviation is 4°?
   d) given that the highs are normally distributed, with standard deviation 4°?
2. How many kids in Lake Wobegon are above average? (Read about Garrison Keillor at publicradio.org.) In our language, what is the probability that a (randomly selected) child has "score" [or something] above expectation if the "scores" are distributed:
   a) uniformly? b) exponentially? c) normally?
   (In each case, no other information is needed.)

d) the approximation

Now we can elaborate what this whole Gauss section is about, the normal approximation to the binomial distribution.

Proposition. As the number n of Bernoulli trials with success probability p grows toward infinity, the binomial density approaches

$$H(x) \;=\; 1/[\sigma\sqrt{2\pi}]\; e^{-([x-\mu]/[\sigma\sqrt{2}])^2},$$

in which $\mu = np$ is the expected value of the binomial distribution and $\sigma = \surd(np[1-p])$ is the corresponding standard deviation.

We can use the density as written to estimate individual probabilities.

Return to the casino and the probability $p(493)$ of 493 WINs in 1000 rolls at craps. There, $\mu = 493$ and $\sigma \;=\; \surd(1000[.493].507) \;\approx\; 15.8$. Use, as is customary, values halfway between integers. Then $p(493)$ is the area under the graph of

$$H_2(w) = 1/[15.8\sqrt{2\pi}]\; e^{-([w-493]/[15.8\sqrt{2}])^2}$$

between $w = 492.5$ and $w = 493.5$.

Near the high point, the graph is practically horizontal. Therefore we may take

$$\text{area} \;\approx\; H_2(493) \times \text{width} \;=\; 1/(15.8\surd[2\pi]) \;\approx\; 0.025.$$

Even where the graph is sloping, over a narrow interval, the height at the midpoint is an excellent approximation to the average height. Hence for example (via scientific calculator)

$$p(450) \;\approx\; H_2(450)\,(450.5 - 449.5) \;=\; 1/[15.8\sqrt{2\pi}]\; e^{-(-43/[15.8\sqrt{2}])^2} \;\approx\; 0.00062.$$

[Compare those results with the plotted probabilities in the chart from subsection (a).

The plotted values are calculations of the exact probabilities,

$$p(493) = \binom{1000}{493}.507^{507}.493^{493} \qquad \text{and} \qquad p(450) = \binom{1000}{450}.507^{550}.493^{450}.$$

There are approximations to those values first developed by de Moivre and refined by James Stirling; read about both at University of St Andrews. It was by studying the approximations that Gauss established the limiting behavior of the binomial distribution.]

For intervals of values—as opposed to individual ones—it makes considerably more sense to return to thinking in standard deviations.

Recall that we postponed figuring the probability of turning a profit on 1000 rolls at the craps table. To make a profit, you need 501 or more WINs. Use $w \geq 500.5$. That is 7.5 WINs above expected value, which is in turn

$$z \;=\; 7.5/15.8 \;\approx\; 0.47 \text{ deviations}$$

above expectation. The probability of $z \geq 0.47$ is 0.3192. (How does that value come from the table?) Over that weekend in Vegas, you are odds-on to lose money.

Now you see why casinos put limits on how much you may bet. They want their small advantage to operate over many small bets.

Show up at a casino in an armored truck and say, “I want to bet this here \$500 million on one roll of the dice.” The bet would give the house an expected profit (the negative of your expectation) of

$$.493(\text{-}\$500\text{M}) + .507(\$500\text{M}) \;=\; \$7 \text{ million},$$

with 0.507 probability of making a profit (and of course 0.493 of losing half a billion). No casino would take such a bet; that would be gambling! However, they would be happy to entertain you while you make 10,000 bets of \$50,000 apiece. Then the casino would have expectation

$$10{,}000 \times [.493(\text{-}\$50\text{K}) + .507(\$50\text{K})] \;=\; \text{the same } \$7 \text{ million},$$

with probability 0.9192 of making a profit. There is an outside chance (0.0228) you will win \$3M or more, but it is equally likely that you will *lose* \$17M or more. (All those come from Exercise 2.)

The normal distribution—along with its approximation to the binomial—is enormously important in statistics, and therefore to the empirical parts of science. Examples include social sciences research (as in polling) and medical research (testing effectiveness of medication). It is indispensable for the study of various kinds of reliability. That includes quality control (deciding what is needed to assure a certain "confidence" that something will survive its use or misuse) and validity of statistical inference (deciding how likely it is that some conclusion reflects reality and not merely chance occurrence).

Exercises VIII.B.5d

1. For 1000 rolls, each with probability 0.493 of WIN, we said that the tail ends of the density plot are too low to graph. Certainly, $p(0)$ through $p(440)$ are all individually small. What about collectively? Find the probabilities of:
   a) 450 or fewer WINs. [Our table won't reach (440 or fewer).]
   b) 540 or more WINs.
2. For 10,000 rolls betting $50,000 each at craps:
   a) Show that the probability of not losing (call it simply WINs ≥ 5,000) is 0.0808.
   b) Show that the probability of profit exceeding $3 million (WINs ≥ 5,030) is 0.0228.
   c) Without calculating: Why is losing $17M or more as likely as winning $3M or more?

Section VIII.C. Number Theory

Turning to number theory puts us in the land of the giants. The contributions of Euler and Gauss defined the subject for the developments that followed.

1. Euler and the Theorems of Fermat

It is tunnel vision to focus entirely on Euler's work on Fermat, but that part of his work is both elementary and interesting.

a) Fermat's Little Theorem

Recall Fermat's result (section VII.A.4f(i)) that prime p divides $(a^{p-1} - 1)$ if it does not divide a. Euler proved it via the following:

Theorem 1. If p is prime and a is natural, then p divides $(a^p - a)$.

[Some call *this* one "Fermat's little theorem." Nobody calls the earlier one "Fermat's big theorem."]

Euler's proof was by induction on a.

The base case is immediate: $1^p - 1$ is certainly divisible by p.

Assume now $k^p - k$ is divisible by p. By the binomial theorem,

$$(k+1)^p - \underline{\underline{(k+1)}} = k^p + \binom{p}{1}k^{p-1} + \ldots + \binom{p}{p-1}k^1 + 1 - \underline{\underline{(k+1)}}$$
$$= pk^{p-1} + \ldots + [p(p-1)\ldots2]/[1(2)\ldots(p-1)]\,k^1 + (k^p - k).$$

All the binomial coefficients are divisible by p (next paragraph), and by assumption so is $(k^p - k)$. Therefore $(k+1)^p - (k+1)$ is divisible by p. That completes the induction.

To see that the coefficients are divisible by p, we may simply observe that each is a fraction whose numerator has factor p and denominator does not. (No denominator factor is divisible by p. Why?) Hence p does not cancel out. For more formal evidence, use the example

$$\binom{p}{3} = [p(p-1)(p-2)]/[1(2)\ldots3] = p\,[(p-1)(p-2)]/[1(2)\ldots3].$$

We know it is an integer. Therefore $[1(2)\ldots3]$ divides $p\,[(p-1)(p-2)]$. Because p divides none of 1, 2, 3, it does not divide their product. (Why?) Hence $[1(2)\ldots3]$ is relatively prime to p. Since it divides $p[(p-1)(p-2)]$ and is prime to p, it has to divide $[(p-1)(p-2)]$ (Exercise III.B.4a:5). The blue fraction is an integer, and the binomial coefficient is a multiple of p.

The little theorem is now easy. No matter what a is, p divides

$$a^p - a \;=\; a(a^{p\text{-}1} - 1).$$

If now p does not divide a, then it has to divide $(a^{p\text{-}1} - 1)$.

Exercises VIII.C.1a

1. Find the remainder upon division by 43 of:
   a) 2^{45} b) 2^{85} c) 2^{125}.

b) sum of squares

The second theorem (section VII.A.4f(ii)) states that if prime p has remainder 1 on division by 4, then there are (unique) natural a and b with

$$p = a^2 + b^2.$$

Fermat did not write a proof. Euler's proof was elementary but complicated (even compared to Fermat's proof for Theorem 3 in section VII.A.4f(iv)); look at it in Wikipedia®. At the same page, you will find an advanced proof by Lagrange.)

Interestingly, Lagrange proved something else about sums of squares. We have seen that not every number is the sum of two squares. Lagrange showed that every number is the sum of *four* squares: Given any natural n, you can find a, b, c, d (some maybe zero) such that

$$n = a^2 + b^2 + c^2 + d^2.$$

There is a nice proof (for which you need the language of Gauss, below) by Matilde Lalín at Université de Montréal.

c) Fermat's primes

The third "theorem" was that every natural number $2^{2^n} + 1$ is prime. Euler settled the conjecture via a distinctly Eulerian approach: He looked at what kind of prime *could* divide one of those numbers.

Suppose p divides $2^{2^4} + 1$. That says $2^{2^4} + 1$ is some multiple ip, and

$$2^{2^4} \;=\; ip - 1.$$

Therefore

$$2^{2^5} \;=\; (2^{2^4})^2 \;=\; (ip-1)^2 \;=\; \underline{jp + 1}. \qquad \text{(What does } j \text{ have to be?)}$$

The power 2^{2^5} has remainder 1 upon division by p.

Let k be the first natural exponent for which 2^k has remainder 1, say

$$2^k \;=\; qp + 1.$$

Then k must divide every exponent, like 2^5, with that property. To see that, use 2^5 as example. Apply the division algorithm:

$$2^5 \;=\; Qk + R, \qquad R \text{ satisfying } 0 \le R < k.$$

Then

$$\begin{aligned} \underline{jp+1} &= 2^{2^5} &&= 2^{Qk+R} \\ &= (2^k)^Q\, 2^R &&= (qp+1)^Q\, 2^R \\ &= (lp+1)\, 2^R. && \qquad \text{(What does } l \text{ have to be?)} \end{aligned}$$

Multiply out and transpose to write

$$2^R \;=\; jp - 2^R lp + 1.$$

That says 2^R has remainder 1 on division by p. Since k is the smallest *positive* such power and $R < k$, we infer R cannot be positive. Therefore $R = 0$; k divides 2^5.

> In fact, k has to *be* 2^5. The latter does not have that many divisors, only
>
> $2^0, \quad 2^1, \quad 2^2, \quad 2^3, \quad 2^4, \quad 2^5.$
>
> None of the first five can be k. Imagine if k were say 2^2, so that 2^{2^2} would have remainder 1:
>
> $$2^{2^2} \quad = \quad mp + 1.$$
>
> In that case, we would have
>
> $$2^{2^4} \quad = \quad (2^{2^2})^{2^{4-2}}$$
> $$= \quad (mp + 1)^4 \quad = \quad Mp + 1.$$
>
> That equality is impossible, because $2^{2^4} = ip - 1$ does not have remainder 1. (What *is* its remainder?) That leaves only the sixth candidate, $k = 2^5$.
>
> We said k has to divide *all* the powers for which 2^{power} has remainder 1. By the little theorem, $p - 1$ is one such power. (To apply the little theorem, we have to know p does not divide 2. How can we be sure?) Therefore $k = 2^5$ divides $p - 1$: $p - 1$ is a multiple of 2^5, and
>
> $$p \;=\; (\text{some multiplier}) \times 2^5 + 1.$$

The effect of all that is to narrow the search for prime divisors of $2^{2^4} + 1$ to primes that are of the form (multiple of 32) +1. The list of such numbers is

33, 65, 97, 129, ….

From there, we may scratch the composites. You can check that the prime survivors are

97, 193, 257, ….

We may also dump "…"; no list of possible divisors can go on forever. In fact, even 257 is already too big. The *first* prime divisor of $2^{2^4} + 1$ has to be no more than its square root, 2^{2^3}+ = 256+. To decide whether $2^{2^4} + 1$ is prime, then, we need only try to divide by 97 and 193. Since

$$2^{16} \;=\; 2^{10}2^6 \;=\; 1024(64),$$

we can easily do the operations by hand to establish that $2^{16} + 1$ is prime.

You can apply the argument above to any of the numbers $2^{2^n} + 1$. For example, the only possible prime divisors of $2^{2^3} + 1$ are those of form (multiple of 2^4) + 1:

17, 33, 49, 65, ….

All of those are disqualified. (Why?) For a better example, choose $2^{2^5} + 1$. Its prime divisors have to look like (multiple of 2^6) + 1:

~~65~~, ~~129~~, 193, 257, ~~321~~, ~~385~~, 449, ~~513~~, 577, 641, ….

Euler, a phenomenal calculator, divided by the surviving five candidates to establish that

$$(2^{2^5} + 1) \;=\; 641 \times 6{,}700{,}417.$$

The fifth "Fermat prime" is not a prime number.

It is still not known whether *any* of the "Fermat primes" with $n > 5$ are prime.

Exercises VIII.C.1c

1. Find the remainder of $(2^{32} + 1)$ upon division by 197. (Hint: Use a simple calculator, starting with $2^{10} = 1024 = 5 \times 197 + 39$.)

d) Fermat's Last Theorem

Euler worked on the Last Theorem and some variations. He proved that

$$x^n + y^n \;=\; z^n$$

does not have integer solutions if $n = 3$ or $n = 5$. Recall that Fermat (section VII.A.4f(iv)) had eliminated $n = 4$. Those cases rule out solutions for any multiple of 3, 4, or 5. However, Euler could not establish the general case.

[I take that as conclusive evidence that Fermat did not have a proof, either. You will see support for that opinion below. I'll do my best to separate fact from editorial.]

In one variation, Euler characterized the integer solutions of

$$x^3 + y^3 + z^3 = w^3.$$

The possibility of negative numbers in that equation allows two forms:

$$3^3 + 4^3 + 5^3 = 6^3,$$

with three positives on one side; or the split solution

$$1^3 + 12^3 = 9^3 + 10^3.$$

(The latter figures in the most popular story about Ramanujan. Read from "Durango Bill" Butler.) In another variation, Euler calculated

$$59^4 + 158^4 = 133^4 + 134^4.$$

From such cubic and quartic examples [I assume], he delivered the opinion that solutions to

$$a^n + b^n + \ldots = z^n$$

require at least n n'th powers on the left. That turned out to be false. See David Murphy's report for a sterling account and for related Diophantine equations.

We have seen that Euler demolished the Fermat-primes conjecture and proved the sum-of-squares theorem (as Fermat did not). For the little theorem, he gave more than proof: He generalized it into a new avenue of study.

> If n is not prime, then a^{n-1} might not have remainder 1 on division by n, even if a is relatively prime to n. Use $n = 10$. Clearly 2^9 cannot have remainder 1; its remainder is even. From
>
> $$3^4 = 8 \times 10 + 1,$$
>
> we have
>
> $$3^9 = (3^4)^2 \times 3 = ([\text{multiple of } 10] + 1) \times 3;$$
>
> the remainder is 3. On the other hand,
>
> $$1^4 = 1, \qquad 3^4 = 81, \qquad 7^4 = (49)^2 = 2401, \qquad 9^4 = (81)^2 = 6561.$$
>
> Those are the four numbers relatively prime to 10, raised to the power of how many there are. Euler showed that this pattern always holds.

Proposition. Let n be a natural number and $\varphi(n)$ the count of naturals from 1 to n that are relatively prime to n. For any a relatively prime to n,

$a^{\varphi(n)}$ has remainder 1 upon division by n.

This really is a generalization of the little theorem. If p is prime, then all of $1, 2, \ldots, p - 1$ are prime to p. Therefore $\varphi(p) = p - 1$. If p does not divide a, then it is prime to a. The proposition implies that if p is prime and does not divide a, then a^{p-1} has remainder 1 on division by p.

It is standard to denote this count, Euler's **totient** function, by φ (Greek lower-case letter *phi*). Euler established numerous properties for φ. For one example, he related $\varphi(n)$ to the prime factorization of n. From the relationship, it follows that φ is **multiplicative**: If m and n are relatively prime, then

$$\varphi(m)\,\varphi(n) = \varphi(mn).$$

(Compare Exercise 1.) Then Euler studied other multiplicative functions. For another example, if you add up $\varphi(i)$ for all the divisors i of n, you find the sum is n (as in Exercise 2). That relation introduced functions defined as sums related to the divisors, like the sum of the divisors or their powers. (What is the sum of the zero'th powers of the divisors of n?)

[To quote the beginning of Section VIII.B.2a, Euler produced "worlds of mathematics." If he could not build a whole new theory around the Last Theorem, I can't conceive that Fermat had a proof.]

Proving the Last Theorem was a laborious slog involving international contributors over 350+ years. You can get an excellent account of this history at St Andrews. For half of it, the only successful attacks were for particular primes. (Why is it possible to limit attention to just prime exponents?) The first general approach was introduced around 1819 by Marie-Sophie Germain (1776-1831). (Wikipedia® has a detailed account of her life, but see the article at Agnes Scott College. At the latter site, you will find a remarkable collection of profiles of women in mathematics.) Numerous contributors followed up her methods, producing proofs for a large class of prime exponents. By 1955, the most important watershed was an extremely advanced proposition called the Taniyama (and others) Conjecture (try Wolfram for a synopsis), whose truth would imply the theorem. Finally in 1995, Andrew Wiles presented a proof, which needed some later modification, of the conjecture. That settled the Last Theorem. (Must see: a very personal interview with Wiles at PBS.)

[I saw a short musical titled "Fermat's Last Tango." It was staged within a little church nestled in the headquarters of Citibank. A church inside a bank is an enormous convenience to those of us who worship money. The performance was a satirical revue in which the ghost of Fermat delighted in mocking Wiles's fits and starts in trying to settle the Last Theorem.

In the lobby, there was an exhibition of Fermatiana. Among its items, my favorite was a form letter. It had been written in mid-twentieth century by the math chairman of a well-known school—maybe it was University of Wisconsin. Evidently, its existence owed to the Ramanujan syndrome (end of section IV.A.4), whereby math departments frequently receive communications of earth-shaking discoveries. The form letter read something like this:

> Dear ___________________,
>
> Thank you for your proof of Fermat's Last Theorem. The first error occurs on
> page ______. This invalidates the proof.
>
> Sincerely, …   ]

Exercises VIII.C.1d

1. Count up relatively prime numbers to show that:
   a) $\varphi(4)\ \varphi(5) = \varphi(20)$.
   b) $\varphi(2)\ \varphi(10) \neq \varphi(20)$. Why is this not a contradiction?
2. a) Write down the divisors of 10, including 1 and 10. Figure out their φ-values, then show that the values sum to 10.
   b) Do the same with the divisors of 20.

2. Gauss

To Gauss, number theory was the "queen of mathematics." His *Disquisitiones Arithmeticae* (*Arithmetical Investigations*, 1801) presented and extended previous discoveries to such extent that we may view it as the foundation of modern number theory, the way Euler's *Introductio* … was foundation for modern analysis. The work of Gauss reached into advanced methods in complex functions; we will examine just two elementary areas.

a) congruences

(i) definition

Gauss created an idea that has connections to the division algorithm, along with the same combination of elementariness and power. The definition is this: Let $n \geq 2$ be a natural number and a and b integers; we say a is **congruent to** b **modulo** n, and write

$$a \equiv b \mod n,$$

if $a - b$ is divisible by n. There is no profit in letting n be negative, $n = 1$ is trivial, and $n = 0$ leads nowhere; that is why we restrict n to 2 and above.

Notice that we may say "a and b are congruent," because the relation is **symmetric**: If a is congruent to b mod n, so that

$$a - b = kn,$$

then

$$b - a = (-k)n$$

and b is congruent to a. (See Exercise 3.)

Every integer is congruent modulo n to its remainder upon division by n. Thus,

$$23 = 2 \times 10 + 3 \qquad \text{gives} \qquad 23 - 3 = 2 \times 10,$$

which says that

$$23 \equiv 3 \mod 10.$$

In the same way,

$$-23 = -3 \times 10 + 7 \qquad \text{(remember that remainders have to be nonnegative)},$$

so that

$$-23 \equiv 7 \mod 10.$$

We will refer to the remainders 0, 1, …, $n - 1$ as the **residues** modulo n. With respect to congruences mod n, they make up a complete set of representatives for all integers.

(ii) arithmetic

Arithmetic with congruences is sometimes called "clock arithmetic," by analogy with the passage of hours on a clock. If the time on a 12-hour clock is now 10:00, then in four hours it will be 2:00, because

$$10 + 4 = 14 \equiv 2 \mod 12.$$

(We will frequently write chains of equalities and congruences.) Similarly, five consecutive four-hour periods lead to $10 + 5 \times 4 \equiv 6 \mod 12$ on the clock.

The most fundamental property of congruences is that they are compatible with addition and multiplication. That is, if $a \equiv b$ and $c \equiv d \mod n$, then

$$a + c \equiv b + d \qquad \text{and} \qquad ac \equiv bd \qquad \mod n.$$

For example,

$$23 \equiv 68 \qquad \text{and} \qquad 17 \equiv 82 \qquad \mod 5.$$

Therefore mod 5,

$$40 = 23 + 17 \equiv 68 + 82 = 150$$

and

$$391 = 23 \times 17 \equiv 68 \times 82 = 5576.$$

The general proof for addition is Exercise 4. For the multiplication part,

$$(a \equiv b \mod n) \text{ means } a - b = kn, \qquad (c \equiv d \mod n) \text{ means } c - d = mn.$$

Hence

$$ac = (kn + b)(mn + d) = (knm + kd + bm)n + bd,$$

and it follows that $ac \equiv bd$.

The compatibility allows us to turn big calculations into reasonably small ones. Consider finding the residue of 39^{41} modulo 43. That power is not something we want to evaluate; even 39^2 is no bargain. But we know $39 \equiv -4$. (Everything in this and the next paragraph will be mod 43. Choosing -4 over 39 is a winner, despite the sign, because it lowers the absolute value.) The compatibility implies that the powers of 39 are congruent to the powers of -4.

> Thus,
>
> $$39^4 \equiv (-4)^4 = 256 \equiv -2.$$
>
> (Here we chose the fourth power because 256 is close to a multiple of 43: $256 = 6 \times 43 - 2$.) Then
>
> $$39^{20} = (39^4)^5 \equiv (-2)^5 = -32 \equiv 11,$$
> $$39^{40} = (39^{20})^2 \equiv 11^2 = 121 \equiv -8,$$
> $$39^{41} = (39^{40})[39] \equiv (-8)[-4] = 32.$$

Exercises VIII.C.2a

1. Calculate the residues:
   a) 2^{100} mod 10
   b) 11^{13} mod 15
   c) 12^{13} mod 14.
2. a) Show that any number is congruent mod 3 to the sum of its (decimal notation) digits.
   b) Show that a number (like 123456) is divisible by 3 iff the sum of its digits is divisible by 3.
   c) Do (a) and (b) with 9 in place of 3.
   d) Is 14710131619222528313437 divisible by 9?
   e) Is there a similarly convenient way to tell (from the digits) whether 14710131619222528313437 is divisible by 11?
3. Show that congruence modulo n has the other properties (aside from symmetry) of an **equivalence relation**:
   a) **Reflexivity**: Always $a \equiv a$.
   b) **Transitivity**: If $a \equiv b$ and $b \equiv c$, then $a \equiv c$.
4. Show that if $a \equiv b$ and $c \equiv d \mod n$, then
   $$a + c \equiv b + d.$$

b) the language of congruences

(i) statements

We have encountered numerous discussions couched in the language of remainders. We can simplify all their statements, and often their proofs, by rendering them in the language of congruences. Thus, in Babylonian times (Exercise II.B.1:1), we saw that no natural number with remainder 3 on division by 4 is the sum of two squares. Now we would say that if $n \equiv 3$ mod 4, then no two squares add up to n (Exercise 1). On the flip side, look at the statement of Fermat's sum-of-squares theorem (section VII.A.4f(ii)). It becomes:

If p is prime and $p \equiv 1$ mod 4, then there exist (unique) natural numbers a and b with $p = a^2 + b^2$.

In the next subsections, we will give similar simplifications and proofs for some other statements.

(ii) inverses

Recall the important theorem (Theorem 1 in section III.B.4) that if a and b are relatively prime, then some integer combination $ia + jb$ equals 1. For example, 16 and 21 are relatively prime, and

$$(4)16 + (-3)21 = 1.$$

From the rearrangement

$(4)16 - 1 = (3)21,$ we judge that $(4)16 \equiv 1 \mod 21;$

similarly,

$(-3)21 - 1 = (-4)16$ implies $(-3)21 \equiv 1 \mod 16.$

Whenever c and d have the property that $cd \equiv 1 \mod n$, we say that c and d **are inverses** mod n. (Clearly the relation is symmetric.) If c and d are inverses mod n, we write $c = d^{-1}$ and $d = c^{-1}$. (Read "*c* inverse," not "*c* to the -1.") We just saw that 16 and 4 are inverses mod 21, and 21 and -3 are inverses mod 16. We will say that 4 is *the* inverse of 16 mod 21, because inverses are "unique" (Exercise 2). In general, the integer-combination theorem implies that if a and b are relatively prime, then each has an inverse modulo the other.

On the other hand, it is impossible for a to have an inverse mod b if the two are *not* relatively prime. Thus, 16 cannot have an inverse mod 30. If there were such an inverse k, then we would have

$1 \equiv (k)16 \mod 30.$

Multiply both sides by

$30/(\text{GCD of 30 and 16}) = 30/2 = 15.$

The result would be a contradiction,

$15[1] \equiv 15[(k)16] = 240k \equiv 0 \mod 30.$

(The reason for choosing 30/(GCD of 30 and 16) is that necessarily

$16 \times (30/\text{GCD of 30 and 16}) = (16/\text{GCD of 30 and 16}) \times 30$

is a multiple of both 16 and 30. Indeed, it is the *least* common multiple; see Exercise III.B.4a:3e.)

Once you have multiplicative inverses, you can define the inverse of multiplication, **division**. If d has inverse d^{-1}, we define the **quotient** a/d as ad^{-1}. Naturally, we call

$1/d = 1d^{-1} = d^{-1}$

the **reciprocal** of d.

By the definition, mod 21 we have

$12/4 = 12(4^{-1}) = 12(16) = 192 \equiv 3.$

That figures, because $4 \times 3 = 12$. We also have

$12/5 = 12(5^{-1}) = 12(-4)$ (Check that $5^{-1} = -4$.)

$= -48 \equiv 15.$

That one looks strange, but is in keeping with $5 \times 15 = 75 \equiv 12$.

Exercises VIII.C.2b(ii)

1. Show that:

   a) If m is even, then $m^2 \equiv 0 \mod 4$.

   b) If m is odd, then $m^2 \equiv 1 \mod 4$.

   c) If $n \equiv 3 \mod 4$, then n is not the sum of any pair of squares.

   d) In a primitive right triangle (integer sides having no common divisor), the hypotenuse cannot be 123,456,003 long.

2. a) Certainly $(4)16 \equiv 1 \mod 21$, but also

   $(25)16 = 400 \equiv 1.$

   In what sense are inverses mod 21 unique?

   b) Show that if $(k)16 \equiv 1$ ("*k* is an inverse of 16") mod 21, then $k \equiv 4$ ("*k* is the same as 4") mod 21.

3. a) Find the inverse of 20 mod 43. b) Evaluate 3/20 mod 43.

(iii) theorems

In this new language, we can restate and re-prove ["reprove" means something else] two old results.

Theorem 1. (The Chinese Remainder Theorem) Suppose k, m, n (and perhaps others) is a finite sequence of *pairwise* relatively prime natural numbers. Then given equally many integers a, b, c (others if appropriate), there exists x such that

$$x \equiv a \bmod k, \quad x \equiv b \bmod m, \quad x \equiv c \bmod n, \quad (\ldots \text{ if any}).$$

(Reread the statement in section IV.B.4 to check that this makes an equivalent one, with an exception. Here, the integers a, b, … are not required to be remainders.)

To prove this one, recall that if k, m, n are pairwise relatively prime, then each one is relatively prime to the product of any of the others (as in Exercise IV.B.4:2d.) Therefore mn has an inverse K mod k, kn has an inverse M mod m, km has an inverse N mod n. The required number is

$$x = aKmn + bkMn + ckmN.$$

This number is congruent to a mod k because every term but the first is a multiple of k. Accordingly,

$$x \equiv aKmn + 0 + 0 \equiv a1 \qquad \text{mod } k.$$

Similarly we deduce $x \equiv b$ mod m, $x \equiv c$ mod n.

Now suppose p is prime. If some integer is not divisible by p, then it must be relatively prime to p. We conclude that every integer that is not a multiple of p has an inverse mod p. That will turn out to be good to know, but it immediately leads us to a favorite.

Theorem 2. (Fermat's Little Theorem) If p is prime and a is not divisible by p, then

$$a^{p-1} \equiv 1 \qquad \text{mod } p.$$

[It is worth a revisit just to see an elegant proof that I heard from my colleague, Jay Jorgenson.]

To prove it, look at the $p - 1$ numbers

$$a, 2a, 3a, \ldots, (p-1)a.$$

First, no two of them are congruent mod p. If two were, say

$$ia \equiv ja \qquad \text{with } i > j,$$

then

$$ia - ja = (i - j)a$$

would be divisible by p. That would mean p must divide either $i - j$ or a. (Reason?) But p does not divide a, by assumption; and it cannot divide $i - j$ either, because $i - j$ is between 1 and $p - 2$. Second, none is congruent to 0; each is a product of two factors p does not divide. Therefore those $p - 1$ numbers span all the nonzero residues: One of them is congruent to 1, another is congruent to 2, …, a last one is congruent to $p - 1$.

We conclude that their product is congruent to

$$1(2)(3)\ldots(p-1) \quad = \quad (p-1)!.$$

Of course, their product actually *is*

$$a(2a)3a\ldots(p-1)a \quad = \quad (p-1)!a^{p-1}.$$

Therefore we have

$$(p-1)!\, a^{p-1} \equiv (p-1)! \qquad \text{mod } p.$$

Now take $(p - 1)!$. It is not divisible by p, because none of its factors is. By the statement preceding the theorem, $(p - 1)!$ must have an inverse K mod p. Multiply the last congruence by K to write

$$K(p-1)!\, a^{p-1} \equiv K(p-1)!.$$

That amounts to $a^{p-1} \equiv 1$.

(iv) order

By Fermat's little theorem,

$$2^6 \equiv 1 \equiv 3^6 \mod 7.$$

Modulo 7, the powers of 3 are

$$3^1 = 3, \quad 3^2 = 9 \equiv 2, \quad 3^3 \equiv 3 \times 2 = 6, \quad 3^4 \equiv 3 \times 6 \equiv 4,$$
$$3^5 \equiv 3 \times 4 \equiv 5, \quad 3^6 \equiv 3 \times 5 \equiv 1;$$

the first power with $3^m \equiv 1$ is $m = 6$. Not so with 2:

$$2^1 = 2, \quad 2^2 = 4, \quad 2^3 = 8 \equiv 1.$$

Whenever there is a positive power m of a such that $a^m \equiv 1 \mod n$, there must exist a *smallest* such power. We call that smallest power the **order** of a mod n. From what we just wrote, the order of 2 mod 7 is 3, and the order of 3 mod 7 is 6.

Notice that since $3^6 \equiv 1 \mod 7$, it follows that

3^5 is the inverse of 3^1, $\quad 3^4 = (3^2)^{-1}, \quad \ldots.$

See Exercise 3a.

Fermat also tell us that

$$2^{42} \equiv 1 \mod 43.$$

Is 42 the order of 2 mod 43? We could try $2^2, 2^3, 2^4, \ldots$. Instead, let us recall an analogous situation, the powers of $i = \sqrt{-1}$.

Recall that those powers cycle through $i, -1, -i, 1, i, \ldots$. Every fourth power equals 1, and $i^k = 1$ iff k is a multiple of 4. That is a fundamental truth about periodic behavior. Anywhere you have a process whose repetitions eventually bring you back to the ground state—in this case, repeated multiplication by i eventually gives $i^k = 1$, so that the next multiple returns to i—then the smallest k satisfying the relation divides the others. (Here, we see that i has order 4, and 4 divides those k for which $i^k = 1$.) Moreover, the key to the proof is always the same: the division algorithm.

Theorem 3. If a has an order mod n, then the order divides all the other k such that $a^k \equiv 1 \mod n$.

Instead of writing a general proof, let us outline the proof using 2 mod 43.

Let m be the order of 2. We also have $2^{42} \equiv 1$. By the division algorithm,

$$42 = qm + r, \qquad 0 \le r \le m - 1.$$

Then

$$1 \equiv 2^{42} = 2^{qm+r} = (2^m)^q\, 2^r \equiv (1)^q\, 2^r.$$

That says r, a number smaller than m, satisfies $2^r \equiv 1$. Since m is the smallest positive power with $2^m \equiv 1$, we infer that r is not positive. Therefore $r = 0$, and m divides 42, illustrating the proof.

In our case,

$$42 = 2 \times 3 \times 7$$

has just eight divisors: 1, 2, 3, 6, 7, 14, 21, 42. Therefore we check

$$2^2 = 4, \quad 2^3 = 8, \quad 2^6 = 64 \equiv 21, \quad 2^7 \equiv 2 \times 21 = 42 \equiv -1,$$

and there the checking stops. Since $2^7 \equiv -1$,

$$2^{14} = (2^7)^2 \equiv (-1)^2 = 1.$$

The order of 2 mod 43 is 14.

Now we reunite with another old friend.

Theorem 4. (Fermat's primes) If p is a prime that divides $2^{2^n} + 1$, then $p \equiv 1 \bmod 2^{n+1}$.

> For partial proof, let p divide
>
> $$2^{2^n} + 1 = 2^{2^n} - (-1).$$
>
> That says
>
> $$2^{2^n} \equiv -1 \bmod p, \qquad \text{so} \qquad 2^{2^{n+1}} = (2^{2^n})^2 \equiv 1.$$
>
> Therefore the order of 2 mod p divides 2^{n+1}. Therefore the order of 2 *is* 2^{n+1} (Exercise 4). We also know that $2^{p-1} \equiv 1$. (At least, we know it if we know that p does not divide 2. What guarantees that?) By Theorem 3, 2^{n+1} must divide $p - 1$. In other words,
>
> $p \equiv 1 \bmod 2^{n+1}$. (Compare Euler's proof in section VII.C.1c.)

Exercises VIII.C.2b

1. a) Find a solution of the congruence system
   $x \equiv 2 \bmod 5$, $\quad x \equiv 3 \bmod 8$, $\quad x \equiv 4 \bmod 9$.
   b) Characterize all the solutions.
2. a) Evaluate the residues of 2^{100} mod 7 and mod 43.
   b) We now know that $2^{32} + 1$ is not divisible by any prime below 641. Find its remainders upon division by 3, 5, 7, 11.
3. a) Does there exist k such that $18^k \equiv 1 \bmod 81$?
   b) What is the order of 3 mod 43?
4. In the proof of Theorem 4, why is 2^{n+1}, and not a smaller number, the order of 2?
5. a) The two-digit decimal numeral 25 has the property that $25^2 = 625$ ends in the same numeral. Are there any others?
   b) Is there a three-digit numeral whose square ends in the same numeral?
   [I learned (a) from Lee Child's fictional character "Jack Reacher."]
6. Show that if p is a prime exceeding 5, then (111…1) (decimal numeral with $p - 1$ digits, all 1) is divisible by p. [That one is from Richard Kasna.]

c) the prime-number theorem

The theorem is a statement about the density—actually, about the sparseness—of prime numbers among the natural numbers. Late in his life, Gauss said that patterns of primes had led him to a conjecture about $\Pi(n)$, the number of primes from 1 to n, some fifty years before (around 1795). About that same time, Adrien-Marie Legendre had made an equivalent conjecture.

Proposition. (**The Prime Number Theorem**) For large n,

$$\Pi(n) \approx n/\log_e n.$$

It is important to understand the sense of the approximation. It is that $\Pi(n)$ is **asymptotic** to the other quantity: If (as Wallis would have said it) n is infinite, then

$$\Pi(n)/[n/\log_e n] = 1.$$

It is not true that the two quantities are nearly equal. If one were $10^{100} + 10^{50}$ and the other 10^{100}, then their ratio would be

$$1 + 10^{-50} = 1.00…01 \text{ (49 zeroes)},$$

but their difference would be inconceivably large. (Compare Exercise 1.)

Exercises VIII.C.2c

1. a) Use a count of primes (and a scientific calculator) to find the ratio and difference
   $\Pi(100)/[100/\log_e 100]$ and $\Pi(100) - [100/\log_e 100]$.
   b) Do the same for 1,000.
2. The distribution of primes has some strange properties. Bertrand's postulate says that you can always find a prime between n and $2n$ (for $n \geq 2$). On the other hand, you can find indefinitely long stretches of natural numbers devoid of primes. Show that the one million consecutive integers
   $(10^6 + 1)! + 2,\qquad (10^6 + 1)! + 3,\qquad \ldots,\qquad (10^6 + 1)! + 10^6 + 1$
   are all composite.

Section VIII.D. Algebra

The result that culminated eighteenth-century algebra is the Fundamental Theorem.

Proposition. (The Fundamental Theorem of Algebra) Every nonconstant polynomial with complex coefficients has a complex root.

To take advantage of it, we first need some results from long before.

1. Polynomials

In previous examples, our polynomials always had integer coefficients. Now we will need to allow division. That need alone would force us to expand the set of candidates to at least the rationals. However, our results apply even to polynomials with complex roots. For that reason, we now allow polynomials to have *complex* coefficients, and their single variable to take complex values.

Remember some definitions and properties. In a polynomial written as a sum, the highest power of the variable actually there (having nonzero coefficient) is the **degree**. The term with that power is the **leading** term, and its coefficient is the **leading coefficient**. Neither definition applies to the **zero polynomial**, the one with fixed value 0; it has no leading term, no degree (as opposed to one with fixed nonzero value, which has **degree 0**). We may say *the* zero polynomial, because it is unique.

Our polynomials are *functions*. Accordingly,

$$f(z) = g(z)$$

means that their values match for all complex z. In that case, f and g have to be the same polynomial: same degree, and corresponding terms having the same coefficients.

Thus, it is impossible to have

$$123z^4 + 456z^7 + 789z^{10} \quad = \quad 212223z^{24} + 121314z^{15} + 31z^{10}$$

for all z. If that were the case, then we would divide by the biggest power z^{24} to say

$$123/z^{20} + 456/z^{17} + 789/z^{14} \quad = \quad 212223 + 121314/z^9 + 31/z^{14}$$

for all z. That statement cannot be true: If z is Wallis's ∞, then the left side is 0 and the right 212223.

By similar reasoning, $H(z) = 0$ forbids H to have any nonzero coefficients.

The sum of two polynomials has the greater of their degrees, unless the leading terms cancel. From

$$(uz^m + \ldots)(vz^n + \ldots) \;=\; uv\,z^{m+n} + \ldots,$$

we see that the degree of a product of polynomials is the sum of the degrees.

Now we are ready to present three results.

Theorem 1. (The Division Algorithm) Suppose $f(z)$ and $g(z)$ are polynomials, g not the zero polynomial. Then there exist a **quotient** (polynomial) $q(z)$ and **remainder** $r(z)$ such that

$$f(z) \;=\; q(z)g(z) + r(z);$$

and q and r are unique if we insist that r be either 0 or of *lower* degree than g.

The analogy to the statement for integers is obvious. It is interesting that "r [is] either 0 or of lower degree than g" takes the place of "$0 \le$ remainder $<$ divisor." We extend the analogy to divisibility: If (and only if) $r(z) = 0$, so that

$$f(z) \;=\; q(z)g(z),$$

then we say that $g(z)$ **divides** $f(z)$ (or g **is a factor of** f, or f **is a multiple of** g).

$$\begin{array}{rrrrr}
 & & 2z^2 & -\ 1/2\,z & -\ 9/4 \\ \cline{2-4}
2z+3 \big) & 4z^3 & +\ 5z^2 & -\ 6z & \\
 & 4z^3 & +\ 6z^2 & & \\ \cline{2-4}
 & & -z^2 & -\ 6z & \\
 & & -z^2 & -\ 3/2\,z & \\ \cline{3-4}
 & & & -9/2\,z & \\
 & & & -9/2\,z & -\ 27/4 \\ \cline{4-5}
 & & & & 27/4
\end{array}$$

Take for example

$$f(z) = 4z^3 + 5z^2 - 6z \quad \text{and} \quad g(z) = 2z + 3.$$

In the text box at right, we see the long division. The green entries show that

$$f(z)/g(z) \;=\; 2z^2 + (-z^2 - 6z)/(2z + 3).$$

The effect of the process, so far, is to reduce the degree of the dividend (the polynomial into which we are dividing). That is the basis for a general proof: It would be proof by induction on the degree of the dividend.

We continue to the orange entries, which show

$$f(z)/g(z) \qquad = \qquad 2z^2 - 1/2\ z + (-9/2\ z)/(2z + 3).$$

As long as the degree of the dividend on the right exceeds or equals the degree of $g(z)$, the division can continue. Finally, the red entries indicate

$$f(z)/g(z) \qquad = \qquad 2z^2 - 1/2\ z - 9/4 + (27/4)/(2z + 3).$$

The long division has produced

$$f(z) \;=\; (2z^2 - 1/2\ z - 9/4)\ g(z) + 27/4. \qquad \text{(Check with Exercise 1.)}$$

As for uniqueness, suppose

$$f(z) \;=\; q(z)g(z) + r(z) \;=\; Q(z)g(z) + R(z),$$

with both $r(z)$ and $R(z)$ either zero or of lower degree than $g(z)$. Rewrite

$$[q(z) - Q(z)]\, g(z) \;=\; R(z) - r(z).$$

If $q(z)$ and $Q(z)$ were not identical, then the product on the left would have at least the degree of $g(z)$, whereas the polynomial on the right has either smaller degree or no degree. Therefore $q(z)$ and $Q(z)$ have to be identical, the left side is the zero polynomial, and $R(z)$ and $r(z)$ have to be identical.

Theorem 2. (The Remainder Theorem) Suppose $f(z)$ is a polynomial and u is a complex number. Then the remainder of $f(z)$ upon division by the linear polynomial $z - u$ is $f(u)$.

From the division algorithm, we have

$$f(z) \;=\; q(z)(z - u) + r(z),$$

where $r(z)$ is zero or has smaller degree than $(z - u)$. That divisor has degree 1. Therefore the remainder is zero or has degree 0; it is a (possibly zero) number v. Thus,

$$f(z) \;=\; q(z)(z - u) + v.$$

That equality has to hold for all z. In particular, it is true for $z = u$. Substitute $z = u$ to see that

$$f(u) \;=\; q(u)(u - u) + v \;=\; v.$$

We have shown that the remainder is the number $f(u)$.

Theorem 3. (The Factor Theorem) The complex number u is a root of the polynomial $f(z)$ iff $(z - u)$ is a factor of $f(z)$.

The factor theorem was doubtless known before Cardano. The proof is almost immediate:

u is a root of $f(z)$ iff $f(u) = 0$ (by definition)
iff the remainder of $f(z)$ on division by $z - u$ is 0 (remainder theorem)
iff $f(z)$ has $z - u$ as a factor.

The theorem guarantees that a polynomial of degree n can have no more than n distinct roots.

Imagine that u is a root of

$$f(z) = 4z^3 + 5z^2 - 6z + 7.$$

By the factor theorem,

$$f(z) = q(z)(z - u),$$

in which $q(z)$ must have degree 2. Suppose now $v \neq u$ is a second root of f. Then

$$0 = f(v) = q(v)(v - u).$$

Since $(v - u)$ is not zero, that forces $q(v) = 0$. The factor theorem applies equally to $q(z)$:

$$q(z) = s(z)(z - v),$$

where now $s(z)$ has degree 1. Once we get to first degree, we know exactly how $s(z)$ has to look:

$$s(z) = az + W = a(z - -W/a).$$

From

$$4z^3 + 5z^2 - 6z + 7 = f(z) = a(z - -W/a)(z - v)(z - u),$$

we conclude that $a = 4$ and $f(z)$ can only be zero for $z = u$, v, or $w = -W/a$.

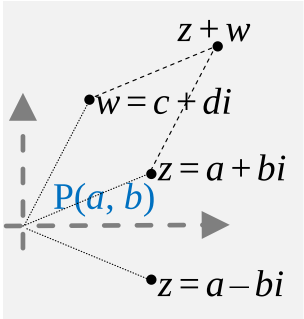
Exercises VIII.D.1

1. Multiply out to verify that
$$4z^3 + 5z^2 - 6z = (2z^2 - 1/2\,z - 9/4)(2z + 3) + 27/4.$$
2. What is the remainder of:
a) $4z^3 + 5z^2 - 6z$ upon division by $z + 1$?
b) $4z^3 + 5z^2 - 6z$ upon division by $2z - 2$? (Reminder: The remainder theorem deals with a specific division.)
c) $z^{43} - 1$ upon division by $z^2 - 1$? (same reminder)
3. a) Guess one solution of the equation
$$z^3 - z^2 - z - 2 = 0.$$
b) Use the solution to factor the cubic.
c) Find all the complex solutions.

2. Complex Numbers

Next we need added understanding of complex numbers. Review as needed the short discussion of their arithmetic in section VI.B.4b.

a) coordinate plane

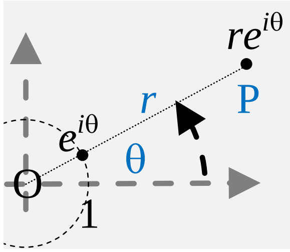


By 1797, there had emerged the picture of the complex numbers in the Cartesian plane. Our usual identification for a point P in the plane, in the figure at right, is an ordered pair (a, b) of real numbers (blue). Instead, let us identify the point with the single complex number $z = a + bi$. In that expression, called the **rectangular form** of the complex number z, we may identify the x-value a as the real part Re z and the y-value b as the imaginary part Im z. The conjugate $\bar{z} = a - bi$ is the mirror-image of z in the x-axis, illustrated in the figure. More important, we may picture complex addition:

$$z + w = (a + bi) + (c + di) = (a + c) + (b + d)i$$

is specified by the **parallelogram rule**, also illustrated.

b) Euler's equation and polar form



Polar coordinates specify the point P(a, b) by the ordered pair (r, θ). There r is distance from the origin and θ is **azimuth**, angle measured counterclockwise from the x-axis to the segment joining P to the origin (provided P ≠ O). Both r and θ are illustrated (blue) at left. The coordinate systems are related by

$$a = r\cos\theta, \qquad b = r\sin\theta.$$

Having agreed that P is $z = a + bi$, we now have

$$z \;=\; r\cos\theta + ir\sin\theta \;=\; r(\cos\theta + i\sin\theta).$$

By Euler's equation, the quantity in parenthesis is $e^{i\theta}$. We then have

$$z = re^{i\theta}.$$

That representation gives the **polar form** of the complex number z.

We will refer to $r = \sqrt{}(a^2 + b^2)$ as the **modulus** of z, denoted by $|z|$. We call θ the **argument** of z, denoted by arg z. The argument is, like the polar angle, not unique; z determines it only to within a multiple of 2π. Notice that $e^{i\theta}$ has geometric meaning. It is, in the last figure, the place where the ray OP crosses the unit circle. Algebraically, since it has polar coordinates (1, θ), it equals

$$1e^{i\theta} \;=\; z/r \;=\; z/|z|.$$

No matter what real number c is, e^{ic} is on the unit circle:

$$\left|e^{ic}\right| \;=\; |\cos c + i\sin c\,| \;=\; \sqrt{}(cos^2\, c + sin^2\, c) \;=\; 1.$$

Finally, $\bar{z}$ has polar coordinates (r, -θ). Consequently $\bar{z} = re^{-i\theta}$, and

$$z\bar{z} \;=\; (re^{i\theta})(re^{-i\theta}) \;=\; r^2e^0 \;=\; |z|^2.$$

c) powers and roots

The product $z\bar{z}$ above is just one example of how polar form facilitates multiplication of complex numbers. Clearly

$$(re^{i\theta})(Re^{ic}) \;=\; (rR)\, e^{i(\theta + c)}.$$

In words, the product's modulus is the product of the moduli, and the product's argument is the *sum* of the arguments. In particular, powers (including rational powers) are given by

$$(re^{i\theta})^k \;=\; r^k\, e^{ik\theta}.$$

Picture those: The powers of a non-real complex number wind around the origin, spiraling outward or inward (depending on r) by steps spanning θ radians. For example,

$$w \;=\; 1 + i\sqrt{}3 \;=\; 2e^{i\pi/3} \qquad\qquad \text{(Verify the second equality!)}$$

has

$$w^2 \;=\; 4e^{i2\pi/3}, \qquad w^3 \;=\; 8e^{i\pi} \;=\; -8, \qquad \ldots, \qquad w^6 \;=\; 64e^{i2\pi} \;=\; 64, \ldots.$$

The powers of w lie along the rays of inclinations 60°, 120°, …, spiraling exponentially away from the origin. If we take instead

$$u \;=\; 1/4 + i\sqrt{}3/4 \;=\; (1/2)e^{i\pi/3},$$

then the powers lie on the same rays but spiral inward toward 0.

Notice that the powers of

$$v \;=\; w/|w| \;=\; e^{i\pi/3}$$

are

$$v^2 \;=\; e^{i2\pi/3}, \qquad v^3 \;=\; e^{i\pi} = -1, \qquad \ldots, \qquad v^6 \;=\; e^{i2\pi} \;=\; 1, \ldots.$$

Those are all on the unit circle, at the vertices of a regular hexagon. Check that

$$v^6 \quad = \quad (v^2)^6 \quad = \quad (v^3)^6 \quad = \quad \ldots \quad = \quad (v^6)^6 \quad = \quad 1.$$

Each of v, v^2, …, v^6 is a sixth root of 1.

The powers of w, coming from multiplication, are uniquely determined by w. The same is not true of roots. The roots of a complex number, as we just saw for $\sqrt[6]{1}$, have multiple values.

In Section VI.B.4c, we left Bombelli needing to calculate

$$\sqrt[3]{-23 + 10i\sqrt{2}} + \sqrt[3]{-23 - 10i\sqrt{2}},$$

which solves the cubic equation

$$x^3 - 27x + 46 = 0.$$

> Check that
>
> $$z = -23 + 10i\sqrt{2}$$
>
> is in the second quadrant and has $|z| = 27$. Therefore
>
> $$z = 27e^{i\theta}, \qquad \text{where } \theta = \cos^{-1}(-23/27).$$
>
> We can no more determine that angle than Bombelli could, but we can approximate:
>
> $$\theta \approx 2.59 \text{ radians}.$$
>
> Then one solution of the equation, corresponding to one cube root of z, is
>
> $$\begin{aligned} x &= \sqrt[3]{z} + \sqrt[3]{\bar{z}} = 3e^{i\theta/3} + 3e^{-i\theta/3} \\ &= 6\cos\theta/3 \qquad\qquad \text{(Why?)} \\ &\approx 3.90. \end{aligned}$$
>
> We can also specify z by
>
> $$z = 27e^{i(\theta \pm 2\pi)}.$$
>
> Those polar forms yield two more solutions:
>
> $$\begin{aligned} x &= \sqrt[3]{z} + \sqrt[3]{\bar{z}} = 3e^{i(\theta + 2\pi)/3} + 3e^{-i(\theta + 2\pi)/3} \\ &= 6\cos(\theta/3 + 2\pi/3) \\ &\approx -5.90; \\ x &= \sqrt[3]{z} + \sqrt[3]{\bar{z}} = 3e^{i(\theta - 2\pi)/3} + 3e^{-i(\theta - 2\pi)/3} \\ &= 6\cos(\theta/3 - 2\pi/3) \\ &\approx 2.00. \end{aligned}$$
>
> Check these solutions against sections VI.B.4c-d.

Exercises VIII.D.2

1. We found six distinct sixth roots of 1 and three distinct cube roots of $-23 + 10i\sqrt{2}$. In each case, are there any others?
2. Evaluate *exactly*, in rectangular form:
   a) all the cube roots of 1;
   b) all the fourth roots of -16;
   c) all the square roots of $7 + 24i$. (Hint: Use the half-angle formulas. Check by squaring.)

---

3. Gauss

The first proof of the Fundamental Theorem came, remarkably, in Gauss's 1799 doctoral dissertation. Boyer says that D'Alembert tried to prove the theorem fifty years before, but without success. Over the years, Gauss published a total of four proofs, based largely on the calculus of complex functions. Our statement of it—ten words, no symbols—is worthy of Gauss. He had a habit of holding his results secret until he could present them in the elegant and polished form he demanded.

a) complete factorization

We wrote that a polynomial with complex coefficients and positive degree must have "a" root. In fact, the Fundamental Theorem implies more.

Theorem 1. A complex polynomial of degree n has exactly n roots, counting multiplicity.

> Suppose $f(z)$ is a complex polynomial. The Theorem says that f must have a complex root r_1. The factor theorem then says that f must have the form
>
> $$f(z) = q_1(z)(z-r_1).$$
>
> The same reasoning applies to q_1, so that
>
> $$f(z) = q_1(z)(z-r_1) = q_2(z)(z-r_2)(z-r_1).$$
>
> The process continues, irrespective of whether $r_1, r_2, \dots$ are different, but it does not continue forever. If f has degree n, then at
>
> $$f(z) = q_n(z)(z-r_1)(z-r_2)\dots(z-r_n),$$
>
> we know $q_n(z)$ has to have degree 0. Its constant value must be the leading coefficient of f.
>
> From the factorization, we have
>
> $$f(r_1) = f(r_2) = \dots = f(r_n) = 0.$$
>
> Moreover, if z is not one of $r_1, \dots, r_n$, then $f(z)$ is the product of nonzero factors. We conclude that every complex polynomial of degree n has precisely n roots, taking multiplicity into account.

We see, further, that every polynomial of positive degree factors completely (aside from the leading coefficient) into its "atoms," the first-degree factors implied by the complex roots.

Exercises VIII.D.3a

1. Suppose the equation
   $$Ax^3 + Bx^2 + Cx + D = 0$$
   has three (not necessarily distinct) complex solutions r, s, t. Show that the sums of their products, one or two or three at a time, are related to the coefficients by:
   a) $r + s + t = -B/A$.
   b) $rs + st + rt = C/A$.
   c) $rst = -D/A$.
2. For the equation and solutions in Exercise (1), show that:
   a) $r^2 + s^2 + t^2 = (B^2 - 2AC)/A^2$
   b) $1/r + 1/s + 1/t = -C/D$.
   Observe that in all five of the relations in Exercises (1) and (2), the solutions play symmetric roles. Compare that with some of Viète's relations in section VI.C.3.)
3. Suppose the equation
   $$Ax^4 + Bx^3 + Cx^2 + Dx + E = 0$$
   has four complex solutions u, v, w, z. Show that:
   a) $uvw + uvz + uwz + vwz = -D/A$.
   b) $uvwz = E/A$.
   c) $1/u + 1/v + 1/w + 1/z = -D/E$.
   Observe again the symmetry. Note further that in (c) here and in 2(b), the sum of the reciprocals of the solutions is the negative of the x-coefficient divided by the constant, regardless of the leading coefficient; and that the same proof would work with a polynomial of any degree.

b) real polynomials

Let us now concentrate on polynomials whose complex coefficients are actually real numbers.

Theorem 1. The non-real complex roots of a polynomial with real coefficients come in conjugate pairs.

Let

$$f(z) \;=\; a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0$$

have real coefficients. Recall the statement, from the arithmetic of complex numbers, that the conjugate of a sum or product is the sum or product, respectively, of the conjugates. Applying the statement repeatedly, we have

$$\overline{f(z)} \quad = \quad \overline{a_n}\bar{z}^{\,n} + \overline{a_{n-1}}\bar{z}^{\,n-1} + \ldots + \overline{a_1}\bar{z} + \overline{a_0}.$$

Because the coefficients are real, we may rewrite

$$\begin{aligned} \overline{f(z)} \quad &= \quad a_n \bar{z}^{\,n} + a_{n-1}\bar{z}^{\,n-1} + \ldots + a_1 \bar{z} + a_0 \\ &= \quad f(\bar{z}). \end{aligned}$$

If now z is a root of f, then

$$f(z) \;=\; 0 \;=\; \bar{0} \;=\; \overline{f(z)} \;=\; f(\bar{z}).$$

In words, if z is a root of a real polynomial, then so is the conjugate $\bar{z}$. That proves Theorem 1.

(i) real quadratics

For a quadratic polynomial $ax^2 + bx + c$ with a, b, c real, the Fundamental Theorem implies two complex roots. By Theorem 1, they have to be both real or both non-real.

Our experience tells us that if the discriminant

$$\Delta = b^2 - 4ac$$

is positive, then the quadratic formula

$$x = (-b \pm \sqrt{}\Delta)/2a$$

names the two different real roots; if $\Delta = 0$, then it names the real double root; and if $\Delta < 0$, then

$$z \;=\; -b/2a + i\sqrt{}-\Delta/(2a) \qquad \text{and} \qquad z \;=\; -b/2a - i\sqrt{}-\Delta/(2a)$$

form the conjugate pair of non-real roots.

(ii) real cubics

For any polynomial of odd degree, the pairing of non-real roots means that there must be an odd number of real roots. In particular, every real polynomial of odd degree must have at least one real root.

Here, let us focus on cubics. We will use our knowledge of roots of complex numbers (section VIII.D.2c) to characterize the sets of solutions of cubic equations. To do so, we have to make an adjustment, based on the multiple values that roots of numbers can take.

To see what is needed, consider our standard cubic equation

$$x^3 + bx + c \;=\; 0.$$

The discriminant is

$$\Delta = c^2 + 4b^3/27.$$

Write

$$A_+ = \sqrt[3]{-c/2 \;+\; \sqrt{\Delta/4}} \qquad \text{and} \qquad A_- = \sqrt[3]{-c/2 \;-\; \sqrt{\Delta/4}}.$$

The first result of Cardano's substitution (section VI.B.2) amounted to

$$x = A_+ + b/3A_+.$$

We turned that form into

$$x = A_+ + A_-.$$

The equivalence of the two forms depended on the relation

$$A_+ A_- \;=\; \sqrt[3]{c^2/4 \;-\; \Delta/4} \;=\; -b/3.$$

The adjustment we have to make is this: For that relation to hold, we must choose for A_- the complex value whose argument is opposite that of A_+. If the arguments do not cancel—equivalently, if they do not add up to a multiple of 2π—then their product $(A_+ A_-)$ cannot be the real number $-b/3$. As long as we adhere to this rule, either form of the Cardano result produces *all* the solutions of the equation.

The easiest case is when $\Delta = 0$, as in

$$x^3 - 27x - 54 = 0.$$

There $A_+ = \sqrt[3]{27}$. If we use the real value $A_+ = 3$, then we need $A_- = 3$. Our choice yields the solution

$$x = 3 + 3.$$

Alternatively, we may use $A_+ = 3e^{i2\pi/3}$. That forces $A_- = 3e^{-i2\pi/3}$. Those are values of equal modulus and arguments ±120°. As a result, their imaginary parts cancel and their real parts sum to $-3/2 + -3/2$. Using the third cube root, $A_+ = 3e^{-i2\pi/3}$, clearly gives the same solution. In that way, $\Delta = 0$ always leads to a single real solution on one side of 0 and a double real solution half as far on the other side.

The second case has $\Delta > 0$, as in

$$x^3 - 27x - 90 = 0.$$

In that one,

$$A_+ = \sqrt[3]{45 + 36} \qquad \text{and} \qquad A_- = \sqrt[3]{45 - 36}.$$

We may add the two real cube roots $\sqrt[3]{81}$ and $\sqrt[3]{9}$ to get one real solution. Instead, we may take

$$A_+ = \sqrt[3]{81}e^{i2\pi/3}, \qquad \text{forcing} \qquad A_- = \sqrt[3]{9}\, e^{-i2\pi/3}.$$

Those have opposite arguments but different moduli. That means their imaginary parts *do not cancel*, and the sum $(A_+ + A_-)$ is a non-real solution. The last choice

$$A_+ = \sqrt[3]{81}e^{-i2\pi/3} \qquad \text{demands} \qquad A_- = \sqrt[3]{9}\, e^{i2\pi/3},$$

producing the conjugate solution. That always happens when $\Delta > 0$: There is one real solution, two non-real (necessarily conjugate) solutions.

For the final case $\Delta < 0$, we have the example at the end of section VII.D.2c,

$$x^3 - 27x + 46 = 0.$$

In that example, we made A_- the conjugate of whichever value we used for

$$A_+ = \sqrt[3]{-23 + 10i\sqrt{2}}.$$

Necessarily, each solution $A_+ + A_-$ was real. But the cube roots are offset from the ±60° and ±120° rays. The offset gives the arguments of the three A_+ candidates different cosines. The resulting sums $A_+ + A_-$ have different real parts. That is why $\Delta < 0$ always results in three unequal real roots.

c) complete factorization of real polynomials

Complete complex factorization leads to a special form for real polynomials.

Theorem 2. Every polynomial with real coefficients is the product of real linear factors and irreducible real quadratic factors.

(According to its title, Gauss's doctoral thesis is aimed at proving this result.)

Real numbers are complex numbers. Therefore a polynomial $f(x)$ with real coefficients, degree n, and leading coefficient a factors into the product

$$f(x) = a(x - z_1)\ldots(x - z_n)$$

of linear factors corresponding to its *complex* roots.

The non-real roots occur in conjugate pairs. Say z_1 is not real, and $z_2 = \overline{z_1}$. Then

$$\begin{aligned}(x - z_1)(x - z_2) &= x^2 - (z_1 + z_2)\,x + z_1 z_2 \\ &= x^2 - (z_1 + \overline{z_1})\,x + z_1\overline{z_1} \\ &= x^2 - (2\,\mathrm{Re}\, z_1)\,x + |z_1|^2.\end{aligned}$$

That product is a quadratic polynomial with *real* coefficients. It cannot be broken into the product of real linear factors, because those would imply real roots, and we already know that its only roots are z_1 and $\overline{z_1}$. Consequently it is **irreducible**: It does not break up into real factors of lower degree.

Multiplying out all other pairs of factors of f having conjugate non-real roots, we see that f decomposes into the product of three types of factors:

its leading coefficient;

some number m (possibly 0, $\leq n$, with the parity of n) of linear factors $x - r$ corresponding to its real roots;

and $(n - m)/2$ irreducible quadratic factors with real coefficients.

Take for example

$$g(x) \quad = \quad x^5 - x^4 + 4x^3 - 4x^2 + 16x - 16.$$

It is fairly clear that

$$g(x) \quad = \quad x^4(x-1) + 4x^2(x-1) + 16(x-1) \quad = \quad (x-1)(x^4 + 4x^2 + 16).$$

(Even without that observation, we can see that $x = 1$ is a root. That makes $(x - 1)$ a factor, whose partner factor we may determine by division.)

The partner has no obvious factors, but it is amenable to the quadratic formula:

$$z^2 \;=\; (-4 \pm \sqrt{[16 - 64]})/2 \;=\; -2 \pm 2i\sqrt{3}.$$

The four square roots of the two numbers on the right give us the other roots of g. Sketch $-2 + 2i\sqrt{3}$ in the plane, to see that

$$-2 + 2i\sqrt{3} \;=\; 4e^{i2\pi/3}.$$

Therefore

$$z^2 \;=\; -2 + 2i\sqrt{3} \qquad \text{yields} \qquad z \;=\; \pm 2e^{i\pi/3} \;=\; \pm(1 + i\sqrt{3}).$$

(Verify that statement by squaring the number on the right.) In a similar way,

$$z^2 \;=\; -2 - 2i\sqrt{3} \qquad \text{yields} \qquad z \;=\; \pm 2e^{-i\pi/3} \;=\; \pm(1 - i\sqrt{3}).$$

In view of those roots, we can factor g as

$$g(x) \quad = \quad (x-1)(x - [1 + i\sqrt{3}])(x + [1 + i\sqrt{3}])(x - [1 - i\sqrt{3}])(x + [1 - i\sqrt{3}])$$

The product of factors #2 and #4 is

$$x^2 - [1 + i\sqrt{3}]x - [1 - i\sqrt{3}]x + [1 + i\sqrt{3}][1 - i\sqrt{3}] \quad = \quad x^2 - 2x + 4.$$

The product of factors #3 and #5 is

$$x^2 + [1 + i\sqrt{3}]x + [1 - i\sqrt{3}]x + [1 + i\sqrt{3}][1 - i\sqrt{3}] \quad = \quad x^2 + 2x + 4.$$

That means

$$g(x) \quad = \quad (x-1)(x^2 - 2x + 4)(x^2 + 2x + 4),$$

with the two quadratic factors irreducible.

Exercises VIII.D.3

1. Prove that every nonzero complex number has exactly n distinct n'th roots.

2. a) Graph the five fifth roots of 32.
b) Use (a) to write the five linear factors of $x^5 - 32$.
c) Use (b) to factor $x^5 - 32$ into linear and quadratic factors having real coefficients (which you may express in terms of trigonometric functions).

3. a) Graph the six complex solutions of
$$x^6 + 64 \;=\; 0.$$
b) Use (a) to write the polynomial as the product of quadratic factors with real coefficients.
c) How come there are no linear real factors?

Section VIII.E. The Astronomers

Before 1600, no earthbound telescope had turned to the sky. By 1801, there existed national and university observatories—plus wealthy individuals—possessed of huge telescopes that had extended the visible universe, along with man's understanding of it.

One of the individuals was the English musician and composer William Herschel (1738-1822). (He was originally Friedrich Wilhelm Herschel. He was born in Germany, like George I. In 1738, George was King of England, and like Herschel, spoke no English.) His characteristics included curiosity. Repeating Newton's experiment in using a prism to separate colors in sunlight, he discovered that the place outside where red fell became heated. He concluded that some sort of invisible energy, relating to red as red relates to orange, was part of the original sunlight. He also had multiple talents, one of them being skill in shaping mirrors for Newtonian reflecting telescopes. (See section VII.B.4d.) He used telescopes (originally bigger than a human, rather than bigger than a building) to hunt for and map double stars. Over decades of observing, he realized that some pairs of stars are "binary systems." That is, they do not simply lie in roughly the same direction as seen from Earth; they are actually orbiting each other, bound necessarily by gravity. He confirmed not only that the stars have proper motion; for pairs and even for systems of more than two, he confirmed that their dance answers to the same law as Kepler's planets and Galileo's cannonballs.

Hunting thus for double stars, Herschel discovered an object too dim to see without optical aid (except maybe for extraordinary eyes) but clearly moving against the background of stars. It moved slowly, as though orbiting out beyond Saturn. A friend used Herschel's observations to confirm that Herschel had discovered Uranus. It was the first planet added to the seven wanderers that had entranced humans for all our time on this one.

Then on January 1, 1801—the first night of the nineteenth century—Giuseppe Piazzi discovered another wandering object. He gave it a name that became "Ceres." By the time Ceres was overrun by the Sun—by the time it became unobservable in evening twilight—Piazzi had managed to observe it for only some weeks. As a result, he could not recover it when the time came for it to become visible west of the Sun, in the dawn. He appealed to the scientific community to invent a method to calculate orbital positions from meager data. The method came, and it yielded a position prediction that recaptured Ceres on the last day of the year. The predicted orbit suggested that Ceres is a small body—the modern estimate of its size is 1/8 Earth—traveling the wide space between the orbits of Mars and Jupiter. It was the first **asteroid**.

The inventor of the method ("of least squares") was Gauss. In that astronomical venture, Gauss was not indulging a hobby. He was a superb astronomer, as his predecessors Euler and Lagrange had been. Officially, it was what he did for a living. It is ironic that the man who earned the nickname "Prince of Mathematicians," whom everybody must have recognized as the third Archimedes, or the second Euler, never held a title like "Professor of Mathematics." The post he entered in 1807, and held for the rest of his life, was in Göttingen's observatory: He was Chief Astronomer.

Chapter IX. The Axiomatization

The nineteenth century brought turmoil to much of the globe. In the early part, France went from an emperor to military defeat and a king imposed by others. Spain lost its South American empire to revolutionaries, who could not keep the whole from breaking into distinct countries (unlike Brazil). In the middle years, the US arrived at the Pacific, then nearly came apart. Japan was unwillingly opened to the world. Russia rose to the status of world power. Germany came together as a county right after the defeat of France, fresh from the Second Empire that followed the Second Republic after the Second Monarchy, in a war that effectively ended in 1945. A year later, Italy became a nation-state. Britannia ruled the waves, but began to grant self-rule to her colonies.

Mathematics underwent something of a revolution as well. The applied side certainly grew, culminating in the complete description of electricity and magnetism. Our interest, however, is on how algebra and the calculus ended up as deductive systems, and on how geometry became a very different system from its Greek origin.

[Much of the material in this chapter is first covered in courses called “advanced calculus” or “introduction to analysis,” or where those are prerequisite. It is not that calculus is needed. What is necessary is enough experience to understand and craft proofs, what the Preface called “a feel for the nature of proof.” Nowadays colleges are increasingly putting in a course specifically designed to ease the transition from algorithmic calculus to proof-based courses; see for example MATH V2000 at Columbia University.]

Section IX.A. Algebra

The Babylonians knew specific verbal prescriptions corresponding to the quadratic formula. From the time al-Khwarizmi gave such equivalents more general form, it was still 600 years before the Italians gave what amounts to a “cubic formula” and “quartic formula.” By such time scales, it is reasonable that 200 years after Bombelli there was still no “quintic formula.” On the other hand, given the pace of mathematical advance from mid-1500’s to mid-1700’s, you would have expected *somebody* to break the logjam. This section is about how answering the quintic question changed the nature of algebra.

1. The Road to Abstraction

a) Lagrange and solution formulas

Recall that Lagrange’s great contributions to the calculus of variations came from generalization. He put existing theory into a general context, much in the manner of Euler and practically simultaneously (section VIII.B.4). Like Euler, he approached problems by making them manifestations of bigger problems, then producing correspondingly big solutions. In that spirit, he chose not to jump into the logjam. Instead, he studied what the successful solution formulas had in common.

In his book of 1770—Berlin days—titled *Réflexions sur la Résolution Algébrique des Équations*, he wrote a theorem about such formulas. Roughly speaking, he said that where there is one, there exists a **resolvent**. The resolvent is a simpler equation, with solutions that lead to those of the original and are in turn given in a special way by *permutations* to the solutions of the original.

This was a genuine jump in abstraction. It made no attempt to solve the equation; rather, he analyzed the *form* of its unknown solutions. In section VI.C.2, we remarked that Viète’s study of relations among the roots similarly dispensed with finding the solutions. Given the involvement of permutations, it appears that Lagrange’s study took off from Viète’s relations. Remember that those are symmetric in the solutions; see some in section VI.C.3 and Exercises VIII.D.3a:1-3. (Accept that those exercises use the

Fundamental Theorem, which was unproven in 1770.) The expressions there have *one* value, unchanged by permutation of the solutions.

His analysis to get to that theorem is out of our league. So are most of the simplifications he then applied. It is elementary, though, to illustrate the situation in our familiar cases. For that purpose, use our elementary notion of **permutation**: rearrangement or reordering. (We will later refine and elaborate considerably on the notion.)

(i) quadratic equations

Recall our usual approach to the quadratic formula.

We wrote the general quadratic equation in the form

$$x^2 + b/a\, x + c/a \;=\; 0,$$

then completed the square to write

$$(x + b/2a)^2 \;=\; (b^2 - 4ac)/4a^2.$$

From there, we got our solutions

$$x_1 = (-b + \surd[b^2 - 4ac])/2a, \qquad x_2 = (-b - \surd[b^2 - 4ac])/2a.$$

(If the discriminant is negative, it does not matter which complex value $\surd[b^2 - 4ac]$ represents.)

Think of it differently. We could have chosen the substitution $t = x + b/2a$ for the purpose of eliminating the degree-(2 – 1)-term.

For that purpose, the substitution works. The quadratic becomes

$$0 \;=\; (t - b/2a)^2 + b/a\,(t - b/2a) + c/a \;=\; t^2 - (b^2 - 4ac)/4a^2. \qquad \text{(Check the algebra.)}$$

The quadratic has yielded to the related equation

$$t^2 \;=\; (b^2 - 4ac)/4a^2,$$

which is simpler, even though it has the same degree.

For the related equation, the two solutions

$$t_1 = \surd(b^2 - 4ac)/2a, \qquad t_2 = -\surd(b^2 - 4ac)/2a$$

lead to the two solutions of the original,

$$x_1 = -b/2a + t_1, \qquad x_2 = -b/2a + t_2.$$

They are in turn given by the original solutions, as

$$t_1 = (x_1 - x_2)/2, \qquad t_2 = (x_2 - x_1)/2.$$

Now write instead $T = t^2$ and consider the resolvent equation

$$T = (b^2 - 4ac)/4a^2.$$

The resolvent has the advantage of lower degree than the quadratic. Its single solution T_1 yields the two solutions of the quadratic by

$$x_1 = -b/2a + \surd T_1, \qquad x_2 = -b/2a - \surd T_1.$$

In turn, that single solution comes out of the quadratic's solutions as

$$T_1 = (x_1 - x_2)^2/2^2. \qquad \text{(Verify.)}$$

Notice that the last expression has a single value under the 2! permutations of x_1 and x_2.

(ii) cubic equations

For the standard cubic, the Italians made the substitution first. Let us use a familiar example with one real solution,

$$x^3 - 27x - 90 \;=\; 0.$$

Substituting $x = u - -27/3u$, we led eventually to

$$x = \sqrt[3]{45+36} + \sqrt[3]{45-36}.$$

From the discussion in section VIII.D.3b(ii), the three solutions of the cubic are

$x_1 = \sqrt[3]{81} + \sqrt[3]{9}$, $x_2 = \sqrt[3]{81}e^{i2\pi/3} + \sqrt[3]{9}e^{-i2\pi/3}$, $x_3 = \sqrt[3]{81}e^{-i2\pi/3} + \sqrt[3]{9}e^{i2\pi/3}$.

The solution process went through the related equation

$$u^6 - 90u^3 + 27^2 = 0.$$

Back in 1545, we stated the obvious, that a sixth-degree equation seems hardly like progress in solving a cubic. But of course, back there we put $v = u^3$, to improve to the lower-degree resolvent

$$v^2 - 90v + 27^2 = 0.$$

The solutions of the resolvent give the solutions of the cubic, and vice-versa.

The resolvent has the solutions

$v_1 = 81$, $v_2 = 9$.

Those two give the three solutions of the cubic, namely the three complex values of either

$\sqrt[3]{v_1} + 9/\sqrt[3]{v_1}$ or $\sqrt[3]{v_2} + 9/\sqrt[3]{v_2}$.

(Do Exercise 1 to match the sets of values.)

To get the resolvent solutions from the cubic solutions, first look at the solutions of the sixth-degree equation. You can verify by substitution that they are

$u_1 = \sqrt[3]{81}$, $u_2 = \sqrt[3]{81}e^{i2\pi/3}$, $u_3 = \sqrt[3]{81}e^{-i2\pi/3}$,
$u_4 = \sqrt[3]{9}$, $u_5 = \sqrt[3]{9}e^{i2\pi/3}$, $u_6 = \sqrt[3]{9}e^{-i2\pi/3}$.

Check next that

$$x_1e^{i2\pi/3} + x_2e^{i4\pi/3} + x_3e^{i6\pi/3} = 3u_5,$$
$$x_3e^{i2\pi/3} + x_2e^{i4\pi/3} + x_1e^{i6\pi/3} = 3u_1,$$

and so on. (You will need the fact that the n n'th roots of 1 always add up to 0; do Exercise 2.) By "and so on," we mean that each of the six u-solutions is given by

$$3u_k = Xe^{i2\pi/3} + Ye^{i4\pi/3} + Ze^{i6\pi/3},$$

in which X, Y, Z is one of the 3! permutations of x_1, x_2, x_3. Finally

$$\begin{aligned} v_1 = 81 = u_1^3 &= (x_3e^{i2\pi/3} + x_2e^{i4\pi/3} + x_1e^{i6\pi/3})^3/3^3 \\ = u_2^3 &= (x_1e^{i2\pi/3} + x_3e^{i4\pi/3} + x_2e^{i6\pi/3})^3/3^3 \\ = u_3^3 &= (x_2e^{i2\pi/3} + x_1e^{i4\pi/3} + x_3e^{i6\pi/3})^3/3^3{}_, \end{aligned}$$

and similarly for $v_2 = 9$ with the other three permutations of x_1, x_2, x_3. (Do Exercise 3 to check the three equations for 9, and it will be clear how to check the three for 81.) In words, the resolvent leads to the solutions of the original cubic, and in turn has solutions given by an expression in x_1, x_2, x_3 that takes on just two values under the 3! permutations of the x's.

(iii) the single form

Always looking for unified (and elegant) theory, Lagrange adjusted the substitutions.

Go back to the quadratic, and change the substitution to $U = 4T$ and the equation to the resolvent

$$U = (b^2 - 4ac)/a^2.$$

Now the lone solution is

$$\begin{aligned} U_1 &= (x_1 - x_2)^2 \\ &= ([-1]x_2 + [1]x_1)^2. \end{aligned}$$

The resolvent's solution is the square of an expression in $x_1$and x_2. The expression is a **root combination**, in which the coefficients are the two square roots of 1; and the combination takes just one value under the 2! permutations of $x_1$and x_2. In terms of U_1, the quadratic's solutions are

$x_1 = -b/2a + \sqrt{}U_1/2$, $x_2 = -b/2a - \sqrt{}U_1/2$.

[“Root combination” is not a standard term, but like “integer combination,” it describes the structure of the expression at hand.]

Next, go back to the cubic and substitute

$U = 3u.$

(It amounts to substituting $x = U/3 - -27/U$ in the original cubic.) The related equation becomes

$U^6/3^6 - 90U^3/3^3 + 27^2 = 0.$

Setting $V = U^3$, we arrive at the resolvent

$V^2 - 90(27)V + 27^4 = 0.$

That one has two solutions,

$V_1 = (27)81, \qquad V_2 = (27)9.$

They lead to the three solutions of the cubic, namely the three complex values of either

$\sqrt[3]{V_1}/3 + 27/\sqrt[3]{V_1}$ or $\sqrt[3]{V_2}/3 + 27/\sqrt[3]{V_2}.$

(Compare Exercise 1.) In turn, the V-solutions are given by

$$\begin{aligned} V_1 &= (3u_1)^3 = (x_3e^{i2\pi/3} + x_2e^{i4\pi/3} + x_1e^{i6\pi/3})^3 \\ &= (3u_2)^3 = (x_1e^{i2\pi/3} + x_3e^{i4\pi/3} + x_2e^{i6\pi/3})^3 \\ &= (3u_3)^3 = (x_2e^{i2\pi/3} + x_1e^{i4\pi/3} + x_3e^{i6\pi/3})^3, \end{aligned}$$

and similarly for V_2 with the other three permutations. The resolvent solutions are the cubes of the root combinations of the original solutions x_1, x_2, x_3; and those cubes assume just two values under the 3! permutations of the three x’s.

(iv) Lagrange’s conclusion

Such was the structure Lagrange established. The formula for degree n gives solutions in terms of the solutions of a resolvent, whose degree is smaller than n and whose solutions are expressions in the n! root combinations

$We^{i2\pi/n} + Ye^{i4\pi/n} + \ldots + Ze^{i2n\pi/n}$

made from permutations $WY\ldots Z$ of the original solutions and the n’th roots $e^{i2k\pi/n}$ of 1.

In the quadratic case, the resolvent had one solution, the lone value assumed by the squares of the 2! root combinations. In the cubic case, the resolvent had two solutions, given by the cubes of the 3! combinations. Similar reduction happens at degree 4: The resolvent has degree 3, with solutions given by an expression that assumes just three values under the 4! permutations of the original solutions. (See Exercise 4 for an example of such an expression.)

For the quintic, Lagrange could not manufacture an expression that (produced the solutions of the quintic and) took fewer than *six* values. That many roots would make the possible resolvents harder to solve than the original. He had proved that the number of values is decisive, but could not establish that no expression with less than five values exists. Unable to find such an expression, he conjectured that no quintic formula exists.

Exercises IX.A.1

1. Verify that the three complex values of either

   $\sqrt[3]{81} + 9/\sqrt[3]{81}$ or $\sqrt[3]{9} + 9/\sqrt[3]{9}$

   are the same as the three solutions of the cubic in subsection (ii).

2. We have seen that $e^{i2\pi/n}$, $e^{i4\pi/n}$, …, $e^{i2n\pi/n} = 1$ are the n distinct n’th roots of 1. Show that their sum is zero. (Hint: Write them as r, r^2, …, r^n.)

3. Verify that
$$\begin{aligned}9 &= (x_3e^{i2\pi/3} + x_2e^{i4\pi/3} + x_1e^{i6\pi/3})^3/27\\ &= (x_1e^{i2\pi/3} + x_3e^{i4\pi/3} + x_2e^{i6\pi/3})^3/27\\ &= (x_2e^{i2\pi/3} + x_1e^{i4\pi/3} + x_3e^{i6\pi/3})^3/27.\end{aligned}$$
(Hint: Verify the first, then notice that the second parenthesis is $e^{i2\pi/3}$ times the first.)
4. a) Find the four solutions of the equation
$$x^4 – 5x^2 + 4 = 0.$$
b) Label the solutions r, s, t, u, in any order. Show that the expression
$$(re^{i\pi/4} + se^{i2\pi/4} + te^{i3\pi/4} + ue^{i4\pi/4})^4$$
has only three values under the 4! permutations of r, s, t, u. (Hint: Do not even *think* of making 4! calculations. Organize, as Euler and Lagrange would. The hint in (3) might help.)
5. Why is n! the number of permutations of n distinct objects?

---

b) Ruffini and the quintic

Paolo Ruffini (1765-1822) was, like Lagrange, Italian. He took up permutations with a more general outlook than Lagrange, who had viewed them strictly as tools fitted to the formula problem. Ruffini studied how sets of permutations are put together. By 1799, he found the connection between the structure of such permutation sets and existence of solution formulas. (It will be easier to describe the connection later.)

As a result, Ruffini proved that there is no quintic formula. In other words, from properties of the permutations of five objects, he deduced that no algebraic expression in the coefficients—no expression applying the arithmetic operations and roots to them—can in all cases evaluate the five complex solutions of a quintic equation. Notice that the same conclusion applies to higher degrees. If there were a sixth-degree formula, then you could apply it to solve
$$0x^6 + Ax^5 + Bx^4 + Cx^3 + Dx^2 + Ex + F = 0.$$

Apparently, his proof was largely not accepted by the mathematical community. Legendre called it “vague.” It was twenty-two years before Augustin Cauchy wrote that he saw Ruffini’s argument as definitive; see Fiona Brunk’s article at St Andrews. Recall that Isaac Barrow’s explanation of the Fundamental Theorem of Calculus (section VII.B.1) met a similar reaction. Maybe Ruffini’s organizational complexity was as mystifying as Barrow’s geometric complexity.

c) Abel

Niels Henrik Abel [obble] (1802-1829) was a Norwegian whose main interest was analysis. Early on, he studied Euler’s limited proof of the binomial series—limited to rational powers—and produced a proof for all real exponents. Later (in his life of 26+ years) he obtained beautiful results on elliptic functions. Those are functions so divorced from the ordinary that they have to be given in terms of what Jesse Douglas called “integrals you can’t do.” (See Boyer’s description.)

Independent of Ruffini, Abel discovered the permutation-to-formula relationship for the fifth degree. Interestingly, Abel’s methods were computational, a step back from the growing abstraction. In his student days in Christiania (now Oslo), Abel had found what he thought was a quintic formula. He spotted a mistake in it, then gave it further analysis. In 1824, he published his discovery.

Many of his results appeared, beginning in 1826, in a mathematics journal started by August Crelle. Those eventually impressed even Legendre and Gauss. But at the time, they did not impress Cauchy. After meeting Crelle in Berlin, Abel visited Paris in the hope of getting a paper accepted by the *Académie des Sciences*. Cauchy assigned evaluation of the paper to reviewers who either did not read it or did not appreciate the depth of Abel's results. (Maybe Cauchy, having accepted Ruffini's proof, considered Abel's to be redundant.)

His papers rejected, Abel was unable to secure an academic post. He was sick with tuberculosis when he returned to Norway. He died there two days before a letter from Crelle arrived, bearing news that a university position in Berlin was Abel's to take.

At the end of the century, when Alfred Nobel funded the prizes that bear his name (How did Nobel get rich?), he did not create a prize in mathematics. The Norwegians wanted immediately to establish a prize honoring Abel's genius, but the separation from Sweden (1905) and subsequent twentieth century crises intervened. The Abel Prize was finally established in 2002 and awarded in 2003, a century after the first Nobels. It is an annual prize, like the Nobels and unlike the Fields Medal, and has gone to a list of longtime contributors to the mathematics of the last fifty years.

d) Cauchy and permutations

Augustin-Louis Cauchy [coe-SHEE] (1789-1857) trained as an engineer. Born with the Revolution, he was actually a staunch royalist. In 1830, the Second Revolution deposed the House of Bourbon, which had been re-imposed with the defeat of the First Empire that had supplanted the First Republic after the overthrow of the monarchy. In protest, Cauchy resigned from the great *Polytechnique* and went voluntarily into exile for eight years. (During that sojourn, one of his posts was Lagrange's old gig at the Turin Military Academy. Lagrange's political outlook had been different, more like "When in Rome…". As a foreigner in both Berlin and Paris, he was careful to observe the norms of his place of residence.)

From the 1820's, Cauchy was a god. That is how papers of the likes of Joseph Fourier (ahead) and Abel became his to judge, even though he did not hold the title of secretary to the *Académie*. For a take on his august standing, read **Ferris** (pages 247-250) about the regard accorded to William Thompson, Lord Kelvin. (The pages relate how Ernest Rutherford had to present a paper contradicting Kelvin's estimate of The Age of the Earth—that is the title of the chapter, pages 231-254—to an audience that included Kelvin.)

Cauchy's innumerable contributions clustered around analysis. We may view him as founder of the calculus of complex numbers, a beautiful theory that is not merely an extension of the real-number calculus. (We mentioned before that Gauss used it in later proofs of the Fundamental Theorem of Algebra.) The theory is beyond our reach, but we will see later some elementary contributions. He also made discoveries in physics, including the theories of light, elasticity, and hydrodynamics. Right here, we focus on his development of permutations.

(i) permutations as functions

Normally we think of permutations as "rearrangements." We call the rearrangement 2 4 1 3 5 a "permutation" of 1 2 3 4 5. Cauchy showed that it is better to think of permutations as functions. To do that, we *define* a **permutation** on a finite set as a one-to-one function with values in the same set.

Thus, the above rearrangement *comes from* the function f defined on the set {1, 2, 3, 4, 5} by

$f(1) = 2,\quad f(2) = 4,\quad f(3) = 1,\quad f(4) = 3,\quad f(5) = 5.$

(Is f actually one-to-one?) We usually think of a function as described by a formula, and we could write one for f:

$f(x) = 2x \qquad \text{modulo } 5,$

with the understanding that if the residue on the right is 0, we use 5 instead. With permutations, though, formulas are typically both inconvenient to write and none too useful.

Using functions, we derive the advantage of a built-in multiplication, namely composition. Cauchy defined the **product** fg of permutations f and g as the composite function given by

$$fg\,(x) = g(f(x)).$$

(It is an unfortunate adjunct of Lagrange's function notation that it conflicts with the way we read, left-to-right. Thus, to apply f first, as we want in $g(f(x))$, we have to write f on the left in fg.) Assuming f is one-to-one, meaning the five values $f(1), \ldots, f(5)$ are different, then the five values $g(f(1)), \ldots, g(f(5))$ are different, provided g is one-to-one. In other words, if f and g are permutations, then so is fg.

This "multiplication" shares with multiplication of numbers the property of associativity, but not commutativity. Composition is always associative: If f, g, and h are functions, then for every x,

$$[fg]h\,(x) \;=\; h(fg\,(x)) \;=\; h(g(f(x)))$$
$$=\; gh\,(f(x)) \;=\; f\,[gh]\,(x).$$

Therefore $[fg]h$ and $f\,[gh]$ are the same function. On the other hand, if F and G are such that

$$F(1) = 2, \qquad G(1) = 1, \qquad \text{and} \qquad G(2) = 3,$$

then

$$F(G(1)) \;=\; F(1) \;=\; 2 \qquad \text{but} \qquad G(F(1)) \;=\; G(2) \;=\; 3.$$

That says GF and FG are different functions.

On a finite set, a one-to-one function is necessarily onto (later discussion), has therefore an inverse. Our first example f makes the assignments

$$1 \to 2, \qquad 2 \to 4, \qquad 3 \to 1, \qquad 4 \to 3, \qquad 5 \to 5.$$

Its inverse f^{-1} simply reverses the arrows. (See Exercise 1.) For each x,

$$f^{-1}(f(x)) \;=\; x \;=\; f(f^{-1}(x)).$$

That means $f^{-1}f \;=\; ff^{-1}$ is the function I given by $I(x) = x$, which we call the **identity** (permutation).

We will denote the set of permutations on $\{1, 2, \ldots, n\}$ by S_n.

(ii) powers and order

Define exponents as we would with numbers:

$$f^1 \text{ means } f, \qquad f^2 \text{ means } ff, \qquad f^3 \text{ means } fff, \qquad \ldots.$$

Notice that, by extension of associativity, no parentheses are needed, even if we go past the third power. Add to those: f^0 means I, and

$$f^{-k} \text{ means } (f^k)^{-1}, \qquad \text{which is the same as} \qquad (f^{-1})^k \qquad \text{(Exercise 2d).}$$

Unlike numbers other than roots of 1, a permutation does not have an infinity of unequal powers.

Take any permutation f in S_n, and look at the list

$$f, \qquad f^2, \qquad f^3, \qquad \ldots.$$

It cannot produce new functions forever; S_n has only finitely many members (Exercise 3). Therefore the list has repetition.

Imagine f^{16} is the first one that matches a previous one, say

$$f^{16} = f^k \qquad \text{(where necessarily } 1 \le k < 16).$$

(Why does there have to be a first one?) In that case, k has to be 1. After all, multiply both sides of the last equation by $(f^{-1})^{k-1}$, and check that the result is

$$f^{16-(k-1)} \;=\; f^{1}.$$

That says $f^{16-(k-1)}$ is already a repeat. Therefore $k - 1$ has to be 0.

Notice in this example that

$$f^{16} = f \qquad \text{forces} \qquad f^{15} = I$$

(compare Exercise 2b), and that 15 must be the *first* power of f that equals I.

We see then that for each f, there exists a lowest power that is I. Its exponent is called the **order** of f. The idea matches that of order in modular arithmetic (section VIII.C.2b(iv)), right down to statements like those in Exercise 4.

If you play card games, then you are familiar with shuffling. Shuffling tries to randomize the order (sequence) of the cards. To do it, you take about half the cards in each hand, then drop some from one hand, drop atop the first ones some from your other hand, drop atop the pile some from the first hand, and so on until all the cards have been dropped into a reordered pile. Players generally agree that it is a fair way to put the cards into an order that does not favor any of them. The "reordered" pile is evidently a permutation on the original sequence. If your shuffle is absolutely consistent—you always take the same number of cards into your left hand (and therefore likewise your right), always drop the same numbers from your two hands in the same sequence of drops—then you are applying the same permutation every shuffle. Accordingly, instead of producing random card orders, you are cycling through a fixed sequence of orders. When the number of shuffles reaches the (mathematical) order of your permutation, you end up with exactly the arrangement you started with.

Exercises IX.A.1d(ii)

1. We defined f on $\{1, 2, 3, 4, 5\}$ by
   $f(x) = 2x$ modulo 5,
   with $f(5) = 5$. Show, without calculating all the values, that the inverse f^{-1} is given (with the same proviso about 5 instead of 0 as a residue) by
   $f^{-1}(x) = 3x$ modulo 5.
2. Prove that for permutations f, g, h in S_n:
   a) The identity acts like 1: $If = fI = f$.
   b) Cancellation applies: If $fg = fh$, then $g = h$.
   c) Inverses are unique: If $fg = I$, then $g = f^{-1}$.
   d) $(f^k)^{-1} = (f^{-1})^k$. (General proof is unnecessary. Use $k = 3$ to give evidence.)
3. a) How many different functions are defined on (all of) $\{1, 2, \ldots, n\}$ and have values there?
   b) How many of those are permutations?
4. Show that for f in S_n:
   a) If $f^k = I$, then the order of f divides k.
   b) One of the powers $f, f^2, f^3, \ldots$ is f^{-1}.

(iii) cycles

Suppose g is a permutation on $\{1, 2, \ldots, n\}$. Start with any k in the set and consider the sequence

$$k,\ g(k),\ g^2(k) = g(g(k)),\ \ldots.$$

It cannot produce new values forever. Somewhere, it starts to repeat. The first repeat has to be k. It could not be, say,

$$g^8(k) = g^2(k).$$

That would say

$$g(g^7(k)) = g(g(k)),$$

which would force the earlier repeat

$$g^7(k) = g(k). \qquad \text{(Why?)}$$

We used this "repeat" argument in (ii), and it is useful enough for us to [how shall we say?] repeat. Let us work an example; take $n = 8$, and let g be the permutation that turns

	1	2	3	4	5	6	7	8
into	3	4	5	6	7	2	1	8.

Suppose we start with 1. The corresponding sequence is

1, $g(1) = 3$, $g(3) = 5$, $g(5) = 7$, $g(7) = 1$.

We will represent those assignments with the **cycle** (1 3 5 7). The cycle is understood to mean that

$1 \to 3 \to 5 \to 7 \to 1$

under g. Notice that we can cycle a cycle:

(1 3 5 7) = (3 5 7 1) = (5 7 1 3) = (7 1 3 5),

because they all make the same functional assignments.

Of the original numbers 1-8, the next unaccounted one is 2. We find the sequence

2, $g(2) = 4$, $g(4) = 6$, $g(6) = 2$.

That means g also performs the cycle (2 4 6). Notice that this second cycle *cannot have any numbers from the first*, because all the numbers in the first cycle are already images of numbers there.

Only the action of g on 8 remains to count. We have the sequence

8, $g(8) = 8$.

That gives the single-element cycle (8), which does not overlap either of the first two.

We now write

$g = (1\ 3\ 5\ 7)(2\ 4\ 6)(8)$.

You can see that we may choose to leave out the (8). The notation suggests multiplication of cycles. The suggestion is exactly right. Our multiplication in S_8 means composition, permutation *followed by* permutation. It is clear that following

$g_1 = (1\ 3\ 5\ 7)$ by $g_2 = (2\ 4\ 6)$ and $g_3 = (8)$

makes the g_1 assignments, then performs the g_2 assignments without disturbing those of g_1, then makes whatever assignments g_3 calls for without changing the previous ones. We may now generalize:

Theorem 1. Every permutation is the product of *disjoint* cycles.

It should be clear that disjoint cycles commute: $g_1 g_2 = g_2 g_1$. In our example, (1 3 5 7) and (2 4 6) operate in separate compartments. Irrespective of which you do first, its results emerge unchanged from the action of the second. We may reorder (permute?) g's constituent cycles. What is more important is that the *set* of constituent cycles is unique.

Theorem 2. The factoring of a permutation into the product of disjoint cycles is unique, except for the order of the cycles.

Suppose you can factor the previous g into disjoint cycles in two ways,

$g = g_1 g_2 g_3 = h_1 h_2 \ldots h_k$.

Then one of the h's must make the assignment $1 \to 3$. That same h must also assign $3 \to 5$, because it is the only one with a 3 in it; the h's are disjoint. Evidently that same one must also assign $5 \to 7 \to 1$, and no other assignments. Hence one of $h_1, h_2, \ldots, h_k$ is g_1. Continuing that way, we conclude that $k = 3$ and each cycle on the right is there on the left. That proves Theorem 2.

[It is standard practice to call $(m_1\ m_2\ \ldots\ m_k)$ a ***k*-cycle**. I love a usage I heard from the late Bernard Vinograde: (1 3 5 7) is a 4-cycle, but (2 4 6) is a “tricycle,” (9 10) a “bicycle,” and (8) a “unicycle.”]

Exercises IX.A.1d(iii)

1. Show that:
   a) The order of a k-cycle is k.
   b) The product
   $g = (1\ 3\ 5\ 7)(2\ 4\ 6)(8)$
   has order 12. More generally, the order of the product $[h_1 h_2 \ldots h_k]$ of disjoint cycles is the least common multiple of the orders of h_1 through h_k.

2. We mentioned the shuffling of cards. Think of a standard deck of 52 cards. At any given moment, call the top card “card #1,” the one right below it “card #2,” and so on to “card #52” at the bottom. In a “perfect (out-)shuffle,” you take cards #1 to #26 in your left hand, #27 to #52 in your right. You drop #52 from your right onto the table, then #26 from your left onto the first drop, then #51 from your right onto the first two, and so on, always alternating right-left. Call the permutation (of the card numbers) so defined *h*.

   a) Show that

   $h(1) = 1$ and $h(52) = 52$.

   b) Find a formula for $h(k)$. It is best to write a two-step formula, one way for $k = 1, \ldots, 26$, a different way for $k = 27, \ldots, 52$.

   c) Find the pair of (unequal) numbers *i* and *j* for which

   $h(i) = j$ and $h(j) = i$.

   Why is there only one pair?

   d) Take any number *m* among the remaining 48—the cards not in (a) or (c)—and write the sequence m, $h(m)$, $h^2(m)$, …. Show that for any of them, the sequence is an 8-cycle.

   e) In view of (a)-(d), what is the order of the perfect out-shuffle (as permutation)?

3. Try the in-shuffle. In that one, you drop from your left hand first, so that #26 goes to the bottom and #27 ends up at the top. The separate formulas are easy to write. Then, if you’re feeling industrious, see that the associated permutation is a single 52-cycle.

(iv) transpositions

The standard name for bicycles is **transpositions**. They play a special role.

Theorem 3. Every permutation is the product of transpositions.

First notice that

$$g_1 = (1\ 3\ 5\ 7) = (1\ 3)(1\ 5)(1\ 7).$$

On the right, the first factor assigns 1 → 3. That assignment is unaffected by the other factors, which do not move 3. The first also assigns 3 → 1, but the second factor moves that resulting 1 to 5, for a subtotal of 3 → 1 . The second further assigns 5 → 1. The combined effect of the first two factors is 1 → 3 → 5 → 1. The third factor assigns that last 1 to 7—which means 5 is carried onward to 7—and also assigns 7 → 1. Therefore the product on the right assigns 1 → 3 → 5 → 7 → 1, same as g_1.

We now conclude that every cycle is a product of transpositions. Since every permutation is a product of cycles, we end up with every permutation as the product of transpositions.

Factoring into transpositions, unlike the disjoint-cycle factoring, is not unique.

Observe that

$$(1\ 3)(1\ 5)(1\ 7) = (1\ 3\ 5\ 7)$$
$$= (3\ 5\ 7\ 1) = (3\ 5)(3\ 7)(3\ 1) \qquad \text{(Exercise 1).}$$

Further, disjoint transpositions commute, like any disjoint cycles, but otherwise order matters. Thus,

$(1\ 2)(1\ 3) = (1\ 2\ 3)$ but $(1\ 3)(1\ 2) = (1\ 3\ 2)$,

and those tricycles are unequal. (Why?)

(v) odd and even

A permutation may factor into the product of different sets of transpositions. However, one thing about the sets has to match: the parity of the number of transpositions.

Theorem 4. In any two factorizations of a permutation into transpositions, the number of factors is even in both, or else odd in both.

We need a sequence of steps to give evidence.

> Work with our previous example *g*, which produces the rearrangement
> 3 4 5 6 7 2 1 8
> Starting with the 3, count how many numbers are now to its right that began to its left (when the numbers were in increasing order). You see that only 2 and 1 are so displaced. Count that as 2 **inversions**. The same is true for the 4, 5, 6, and 7; they give an additional 8 inversions. For the 2, the number 1 is misplaced on the right, the number 8 is correctly placed on the right. That adds 1 inversion. For the 1, there are 0 inversions. The green numbers add up to 2 + 8 + 1 + 0 = 11.

We said *g* is the product of transpositions. Clearly transpositions create inversions. The first transposition factor created some inversions, the second one added *or subtracted* some, and so on. That 11 is the total of inversions given/taken by the transpositions whose product *g* is. We should therefore ask, how many inversions does one transposition factor create or destroy? The answer is that a single transposition factor adds an odd integer of inversions to those of the previous factors.

> Take the transposition (2 6). Applied to the original arrangement
> 1 2 3 4 5 6 7 8,
> it actually switches the 2 and the 6. Moving the 6 to where the 2 *was* puts the intervening 3, 4, 5 on the wrong side of 6, creating 3 inversions. Moving the 2 to where the 6 was puts the 2 on the wrong side of the same intervening 3, 4, 5, adding 3 more inversions. But it also put the 2 rightward of the 6. That makes (3 + 3 + 1) added inversions.
>
> Applied instead to a later rearrangement
> ? *M* *a* *b* *c* *N* ?? ???,
> (2 6) switches the *M* and the *N*, no matter what they are. It produces the arrangement
> ? *N* *a* *b* *c* *M* ?? ???.
> For inversions, the numbers represented by question marks are irrelevant. If they were on the correct side of either *M* or *N* before we applied (2 6), then they stay on the correct side; if not, then they stay incorrect. The ones that (literally) count are *a*, *b*, *c*. In the case where *a* is smaller than both *M* and *N*, before the switch it was on the correct side of one and wrong side of the other, and the same is true after the switch. In this case, the transposition neither added nor lost inversions attributable to *a*. The same reasoning applies in the case where *a* is bigger than both *M* and *N*. In the remaining case, *a* is between *M* and *N* (in numerical value). Then, either it is on the correct side of both before the switch and wrong side of both after, adding 2 inversions; or vice-versa, adding -2. Thus, when we apply (2 6), *a* contributes 0, 2, or -2 inversions to the resulting rearrangement. The same applies to *b* and *c*. That leaves just one other source of contributions. If *N* was on the correct side of *M* before (2 6), then it moved to the wrong side, and vice-versa. The shift of those two contributed either 1 or -1 inversion. The application of (2 6) added to the number of inversions some 0's or 2's or -2's, together with *exactly one* ±1. Of necessity, it added an odd number, possibly negative, of inversions.

Put the steps together now. Our *g* factors into transpositions. The first factor lends an odd number of inversions to the starting arrangement. The second adds or subtracts another odd number, for an even total. The third adds or subtracts an odd number, …. If the number of transposition factors is odd, then the number of inversions in the rearrangement produced by *g* is odd; if the number of factors is even, then so is the number of inversions. Because our chosen *g* produces a rearrangement with 11 inversions, *every* factorization of *g* into transpositions must have an odd number of factors.

We call a permutation **odd** or **even** according to the parity of its number of transposition factors.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	▨

There are many contexts in which one configuration is incompatible with another and the conflict is explainable in terms of odd-even. An unexpected such context is the **15-puzzle**. The puzzle has fifteen square tiles, numbered 1-15, shaded blue at left. They are placed into a square frame four times as wide. You can slide any tile adjoining the remaining empty place (gray) into the hole. Starting from the “original” position shown at left, for example, you can slide the 15 rightward into the hole, or instead slide the 12 down. With either of those moves, the hole slides correspondingly left or up. We ask, is it possible through a series of such moves to produce the position shown at right?

▨	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

The answer is no. Assign the value 16 to the hole, so that the original position is

1 2 … 15 16.

The first move you make results in either

1 2 … 11 12 13 14 16 15

or 1 2 … 11 16 13 14 15 12.

Those rearrangements result from applying the transposition (15 16) or (12 16). Every move you make applies a transposition. Therefore every time you move, you change the resulting product of transpositions (the permutation of the tile numbers) from even to odd, or vice-versa.

At the same time, with every move, you change the position of the hole by one row or one column. At the start, the hole is at

(row, column) = (R, C) = (4, 4),

where $R + C$ is even. Your first move puts it at

(R, C) = (4, 3) or (R, C) = (3, 4),

at either of which $R + C$ is odd. Similarly, every move switches the parity of $R + C$. Consequently every time you move a tile (and therefore the hole), the resulting rearrangement is a permutation of the original *of the same parity* as the resulting $R + C$.

Now look at the proposed rearrangement,

16 1 2 … 14 15.

Its only inversions are for the 16; there are fifteen numbers to its right that started on its left. The needed permutation is odd. But in this rearrangement, $R + C = 1 + 1$. No series of moves can produce simultaneously an odd permutation of the original and an even $R + C$.

Exercises IX.A.1d(v)

1. Trace the images in (3 5)(3 7)(3 1) to show that it equals (3 5 7 1) = (1 3 5 7).
2. Show that an n-cycle is an even permutation exactly if n is odd.
3. Start with four arrows pointing upward and three down, as shown.

   ↑ ↑ ↑ ↑ ↓ ↓ ↓.

   In one “move,” you turn any *two* upside-down, reversing their orientations. Is there any number of such moves that will make them all point upward?

2. Abstract Algebra

a) Galois and groups

Modern mathematics is divided, like all of Gaul, into three parts. They are not neatly separated; in fact, much beautiful mathematics has grown in their overlaps. Still, they are useful categories. Modern **analysis** is what came from the invention of the calculus, although it makes sense to trace it to what

Euler revealed about its power. Modern **topology** studies form and spatial relationships, and as such evolved from geometry. As an axiomatic system, it was created around 1920; look up Felix Hausdorff and Kazimierz Kuratowski. Modern **algebra** studies abstract algebraic structures, and was born the night before Galois was shot.

Évariste Galois (1809-1830)—he did not make it to his twenty-first birthday—was a hothead. He was a radical republican, much as Cauchy was the opposite, in a time and place where anti-monarchical activism was a dangerous pursuit. He was twice imprisoned for political agitating. Politics aside, he was impetuous. After rejections from the *Polytechnique*, he settled for the *École Normale*. He was expelled from it. Finally, he got into some dispute—over a woman, some say—and challenged an ex-friend to a duel. Night before the duel, he had a bad feeling. He started desperately writing a letter, in which he elaborated earlier works (some of which Cauchy had managed *also* to misdirect, like Abel's, even though they cited Cauchy's work on permutations), revealed undeveloped material, and lamented that so little time remained. He ended with a request that his discoveries be presented to Gauss and Carl Jacobi for judgment, not of correctness (whereof he had no doubt), but of importance. Somewhere in there, he introduced the world to what he called *groupes*. It was his valedictory. The bad feeling had good reason: In the duel, he was mortally wounded; he died the next day. It was more than ten years before the letter came to someone who could understand and appreciate it; that was Joseph Liouville, who published it in in his *Journal de Mathematiques Pures et Appliquées* in 1846.

(i) groups

"Abstract algebraic structures" are made up of elements generally unrelated to numbers that nevertheless have number-like properties. [At the time of Bombelli and the development of the arithmetic of complex numbers (section VI.B.4b), the set of complex numbers would have qualified as an abstract algebraic structure. To Cardano, they were not merely abstract; he called them "sophistic."]

The structure that came from Galois's work was the **group**, a combination of constituents satisfying four requirements. We will refer to the requirements as the **axioms of group theory**.

Axiom 1. There is a set G, in which an operation is defined.

An **operation** is a way of combining two elements of the set to yield a third. The archetype operations are (the usual) addition and multiplication in the set of natural numbers. Given two (not necessarily different) naturals a and b, $a + b$ is another one, as is $a \times b$. Subtraction and division are *not* operations: The subtraction $a - b$ does not always yield another natural number; and the same is true of the division a/b. Sometimes, the last sentence is worded as, "Subtraction and division are not *closed* operations." We will avoid that usage. To us, if it is not closed, then it is not an operation.

> You can invent an infinity of operations. Check that each of the following describes a process that produces a natural number from two others:
> a) Define $a \wedge b$ to mean a^b, normal exponentiation.
> b) Define a PSUM b to mean $ab + a + b$, the last expression having the familiar meaning.
> c) Define a WEAVE b as follows: Write the first digit of a, the first digit of b, the second digit of a, the second digit of b, …, until one of them runs out of digits, then continue with the remaining digits of the other one. The two have to be in standard decimal notation—123, not 00123. Thus,
> 12 WEAVE 34 = 1324,
> 1234 WEAVE 56 = 152634.
> d) Define a ONLY b to mean b. Thus, (5 ONLY 6) = 6 and (78 ONLY 34) = 34.

Axiom 2. The operation must be **associative**.

Use (a \$ b) to signify the result of operating a on b. The axiom demands for every a, b, and c in G,

$$(a \$ b) \$ c = a \$ (b \$ c).$$

Addition and multiplication of naturals are two associative operations. (One advantage of using \$ is that it is neutral, not suggestive of addition or multiplication specifically.) Exponentiation is not:

$$(2 \wedge 3) \wedge 4 \ = \ 8^4, \qquad 2 \wedge (3 \wedge 4) \ = \ 2^{81},$$

unequal results. Associativity is not automatic. (Decide about the other examples in Exercise 1a.)

Axiom 3. The operation must have an **identity** element, a member E such that for all a in G,

$$a \mathbin{\$} E \ = \ E \mathbin{\$} a \ = \ a.$$

You should see that requiring (a \$ E) = (E \$ a) is not redundant. No axiom requires, and nothing so far indicates, that order is irrelevant. The natural number 1 is a "left identity" for ONLY, meaning (1 ONLY a) = a, but it is not a "right identity," because (5 ONLY 1) ≠ 5. For exponentiation, 1 is a right identity but not a left identity:

$$a \wedge 1 = a, \qquad \text{but} \qquad 1 \wedge 2 \neq 2.$$

Multiplication has identity 1, because $a \times 1 = 1 \times a = a$, but addition does not have an identity, because no natural E satisfies $b + E = b$.

Axiom 3 demands that there be at least one special element within G. Thereby, it excludes the possibility that G be the empty set. But the axiom happens to guarantee that there is *only* one identity.

Theorem 1. A group's identity is unique.

Suppose E and F are elements that satisfy

$$a \mathbin{\$} E \ = \ E \mathbin{\$} a \ = \ a \quad \text{and} \qquad a \mathbin{\$} F \ = \ F \mathbin{\$} a \ = \ a$$

for every a. Then in particular, E works on the left with F, meaning $E \mathbin{\$} F = F$; and F works on the right with E, meaning $E \mathbin{\$} F = E$. It follows that E and F are two names for one element,

$$E \ = \ E \mathbin{\$} F \ = \ F.$$

We may therefore speak of *the* identity in a group.

Axiom 4. Every element of G must have an **inverse**: If a is any element of G, then there must exist an element denoted by a^{-1} (again "a inverse," not "a to the -1") with

$$a \mathbin{\$} a^{-1} \ = \ a^{-1} \mathbin{\$} a \ = \ E.$$

The conception of 0 and the negative numbers, together with Brahmagupta's extension of arithmetic to them (section IV.A.3), created the set **Z** of integers. You can check that under the operation of addition, **Z** satisfies the four axioms. It is a group. The set $\mathbf{Q}^+$ of *positive* rational numbers is a group under the operation of multiplication. (Verify both in Exercise 2.)

The theory born out of just these four axioms is vast. We will add a scant two elementary results to the humble beginning of Theorem 1.

Theorem 2. In a group:

a) Cancellation applies, on either side. That is, if either

$$a \mathbin{\$} b = a \mathbin{\$} c \qquad \text{or} \qquad b \mathbin{\$} d = c \mathbin{\$} d,$$

then necessarily $b = c$.

b) Inverses are unique (Exercise 3a).

To argue (a), assume $a \mathbin{\$} b = a \mathbin{\$} c$. Apply the element a^{-1}, whose existence is guaranteed by Axiom 4, *on the left* to both. The definition of "operation" requires equal results:

$$a^{-1} \mathbin{\$} (a \mathbin{\$} b) \ = \ a^{-1} \mathbin{\$} (a \mathbin{\$} c).$$

By Axiom 2, we may replace the left and right sides by

$$(a^{-1} \mathbin{\$} a) \mathbin{\$} b \ = \ (a^{-1} \mathbin{\$} a) \mathbin{\$} c.$$

By Axiom 4, those two are

$$E \mathbin{\$} b = E \mathbin{\$} c.$$

By Axiom 3, that last equation says $b = c$. We can work similarly to cancel on the right.

Notice that the above argument relies on just the four axioms and the nature of operations.

Exercises IX.A.2a(i)

1. a) Show that PSUM and ONLY are associative, but WEAVE is not.
   b) Prove that PSUM does not have an identity.
2. Verify that **Z** with addition and $\mathbf{Q}^+$ under multiplication are groups.
3. Show that in a group:
   a) The inverse a^{-1} is unique: $a \$ b = E$ forces $b = a^{-1}$. Hence we may say *the* inverse of a.
   b) The inverse of the inverse of a is a.
   c) The inverse of a "product" is the product of the inverses, but reversed:
   $(a \$ b)^{-1} \ = \ b^{-1} \$ a^{-1}$.
4. a) Show that if a is an element in a finite group, then the first element repeated in the list
   $a, \qquad a^2 = a \$ a, \qquad\qquad a^3 = a \$ a \$ a, \ldots$
   of its "powers" is a. (Remember that by associativity, no parentheses are needed.)
   b) Give an example of a group in which that same list might fail to repeat any element.

(ii) two finite examples

We will look at two specific instances of more general examples. In these two, we will begin to see how groups can have different structures.

+	**0**	**1**	**2**	**3**	**4**	**5**
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	**2**	3	4

The first group uses addition modulo 6. Take the set

$\mathbf{Z}_6 = \{0, 1, 2, 3, 4, 5\}$

of residues modulo 6, and define the operation "+" by

$a + b$ = the mod 6 residue of the integer sum of a and b.

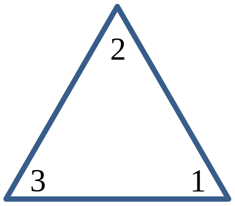
Check that the table at right accurately tabulates the results. The table clearly reflects an operation, because all the results are in $\mathbf{Z}_6$. There is clearly an identity, 0, since the row and column headed 0 have the same numbers as the (shaded) headers. There are inverses, because every row and column has a 0 in it. Only associativity is not evident. To verify it, we would need to compare 6 × 6 × 6 results with an equal number of others. However, if we accept that addition within the integers is associative and that congruence is compatible with that addition, then associativity is inherited.

In $\mathbf{Z}_6$, the operation is commutative. That is easy to see, because the table is symmetric about the main (\) diagonal. A commutative operation is special; when a group has one, we say the group is **abelian** (honoring Abel's studies). It is customary to call the operation in an abelian group "addition," even if it is very different from ordinary addition. (Remember the example of $\mathbf{Q}^+$, Exercise 2 above.) If you adhere to that practice, then you denote its identity by "0" and the inverse of a by "$-a$."

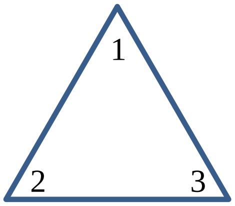
The second example is S_3. Cauchy's analysis made it clear that S_n is always a group: There is multiplication, it is associative, I has the identity property, and every permutation has an inverse. If $n \geq 3$, then S_n is not abelian. In that case, S_n includes (1 2) and (1 3), and we have seen that

$(1\ 2)(1\ 3) \ \neq \ (1\ 3)(1\ 2)$.

Choosing S_3 gives us a group with the same number of elements as $\mathbf{Z}_6$, but differing in a fundamental way. It also gives us a chance to see why S_n is called the **symmetric group** on n elements (which explains the letter "S").

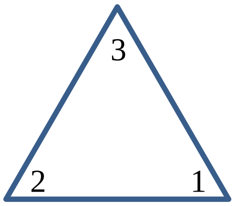


Let us study the set of **symmetries of the triangle**. Cut an equilateral triangle (as nearly as you can) from a sheet of paper. Write "1" on front and back of the corner now on your right, as in the figure at left. Similarly, write "2" front and back at the top, "3" at the remaining corner.



A **symmetry** is a transformation that leaves the paper triangle occupying the same space. For example, the transformation ROT120 that rotates the triangle counterclockwise through 120° about its center (the meeting of the medians) brings the triangle to the position shown at right. The vertices change place, but the triangle occupies the previous space. Similarly, we can apply ROT240, ROT360, ….

> You can see that the list actually has just three symmetries,
>
> ROT120 = ROT(120 ± any multiple of 360°),
> ROT240 = ROT(240 ± any multiple of 360°),
> ROT360 = ROT(0) = ROT(± any multiple of 360°).
>
> There are exactly three others. If we flip the triangle about the median from the vertex on the right, then the original position flips to the one at right. Call that transformation FLPRT. Notice that it *is* new: The rotations leave the corners oriented 1-2-3 counterclockwise, whereas FLPRT leaves them clockwise. Similarly we define flips FLPTOP about the vertical median and FLPLF about the median from the left corner. (Do Exercise 1 to convince yourself that there are no other symmetries of the triangle.)



Now define a multiplication of these symmetries by **composing**, what we did to form the composite $g(f(x))$ of functions f and g. In simpler words, we will follow one symmetry by a second. When we do that, the result is another symmetry. Use your triangle to check that

(ROT120)(ROT240) = ROT360 = ROT(0),
(FLPRT)(FLPTOP) = ROT240,
(FLPLF)(ROT120) = FLPTOP.

Because the result is a symmetry, "compose" or "follow" defines an operation on the set of symmetries.

As usual, associativity is laborious actually to establish. Let us simply agree that composition of transformations is associative, as it is with functions. There is clearly an identity, because ROT(0) preceding or following another symmetry leaves the latter unchanged. Finally, ROT120 is the inverse of ROT240, and ROT(0) and all three flips are their own inverses. The symmetries constitute a group. In fact, the group is indistinguishable from S_3 (Exercise 2), so we will denote it by S_3.

Groups are certainly abstract structures, but their connection to symmetry has yielded concrete results in the sciences. Crystallography is one area in which symmetry has been applied to explain, and more importantly to *predict*, properties of materials. Separately, that predictive ability underlies explorations in subatomic physics, where the discovery of a particle with one set of characteristics (mass, charge, spin) leads to a search for another with mirror-image properties. For a last example: Our "flips," chemistry, and the physics of light come together in the study optical isomers.

Exercises IX.A.2a(ii)

1. a) Without enumerating them, show that there are just six symmetries of the triangle.
   b) How many symmetries are there of the square? of the regular *n*-gon? (Hint for both: Decide where vertex #1 goes; then you have exactly one more decision to make.)

2. Show that each symmetry of the triangle effects a permutation of the vertices we named #1, #2, and #3. Then give three examples to show that following one symmetry by another effects the permutation that is the product of the corresponding permutations.
3. Show that the group of symmetries of the triangle is not abelian.
4. We saw (Exercise 2) that the symmetries of the triangle match the permutations in S_3. Is it always true that the symmetries of the regular n-gon match the permutations in S_n?
5. a) Exhibit two similarities and two differences between the way the operations work, in the group **Z** of integers under addition, versus the group $\mathbf{Z}_6$ under addition mod 6.
b) Exhibit one similarity and one difference between the group **Z** of integers and the group **Q** of rational numbers, both under addition.

(iii) subgroups

We will see that everything we need about groups is found in the symmetric groups. We stay with unspecified groups for this subsection, but henceforth we restrict our attention to *finite* groups.

Let G be a group, and call its operation *multiplication*. We will use normal algebraic notation and write (g multiplied by h) as gh. It would be natural to call the identity “1,” but we need that symbol for counting; we borrow from permutations and write I.

Take an element g from G. Look at its powers

$$g, g^2, g^3, \ldots.$$

By the “repeat” argument, there is $k + 1 \geq 2$ for which

$$g^{k+1} = g$$

is the first repetition on the list. (Remember that G is finite.) In that case, $g^k = I$, and k is the first exponent with that property. We say k is the **order** of g. That means

$$g, g^2, \ldots, g^k = I$$

are all distinct. It also means $g^{k-1} = g^{-1}$. (Compare Exercise 1.)

The subset of G

$$H = \{g, g^2, \ldots, g^k = I\}$$

is itself a group under G’s multiplication. Multiplication is an operation in H, because H is closed:

$$g^i g^j = g^{(i+j \text{ modulo } k)},$$

with the convention that g^0 means I. Multiplication is associative, because we are in a group G. We just said that H has I in it, and it has inverses:

$$(g^j)^{-1} = g^{k-j} \qquad \text{(Exercise 2).}$$

Therefore H meets the four requirements.

When a subset J of a group G is itself a group under G’s operation, we say J is a **subgroup** of G. The subgroup of powers of the element g is called the **cyclic group of** (more formally, **generated by**) g. If it occupies all of G, then we call G a **cyclic group**.

Return to the subgroup H of powers of g. See that the key statement above was that H is closed under the multiplication in G. In a finite group, that by itself guarantees that a subset is a subgroup.

Theorem 3. If the nonempty subset J of G is closed under G’s operation, then J is a subgroup of G. (Why does J have to be nonempty?)

Under the assumption, J has an associative operation. Take any j in J. All the powers of j are in J, because J is closed. We know that I and j^{-1} are among those powers. Therefore J has identity and inverses. It is a subgroup of G.

Just as an integer has automatic divisors, itself and 1, so a group G has automatic subgroups, G itself and $\{I\}$. Check that both are closed under G's multiplication. [Recall that the math word for "automatic" is "trivial." If you adopt it, then you have to say of a group with only the automatic ones that it has "no nontrivial subgroups." I will take the liberty I did with divisors and say it has "no subgroups."] In general, there will also be other subgroups. (Compare Exercise 4.)

In $\mathbf{Z}_6$, there are two others. (Compare Exercise 5.)

A subgroup can hold just 0. (Remember the operation; see Exercise 3.) If additionally it holds 1, then it has to hold also

$$1 + 1 = 2, \qquad 1 + 1 + 1 = 3, \qquad \ldots, \qquad 1 + 1 + 1 + 1 + 1 = 5$$

along with 0. That would fill all of $\mathbf{Z}_6$ (which we conclude is cyclic). A similar thing happens if it holds 5. (Try it.) Those are the trivial possibilities.

If our subgroup has neither 1 nor 5, but does have either 2 or 4, then it has

$$2 + 2 = 4 \text{ and } 2 + 2 + 2 = 0 \qquad \text{or} \qquad 4 + 4 = 2 \text{ and } 4 + 4 + 4 = 0.$$

Either way, it is $\{2, 4, 0\}$. If it holds none of 1, 2, 4, 5, then it has to be $\{3, 0\}$. Those are the others.

One way to make the search for subgroups systematic is to count their elements. The number of elements of a group or a subgroup is its **order**. Notice that this usage or "order" does not conflict with the earlier. If the order of element g is k, then

$$g, \qquad g^2, \qquad \ldots \qquad , \qquad g^k = I$$

are the distinct elements of g's subgroup; the order of g is the order of its subgroup.

What is generally considered the most important theorem for finite groups came right out of Lagrange's pioneering fiddling with permutations.

Theorem 4. (Lagrange's Theorem) The order of a subgroup divides the order of the group.

We will illustrate the general argument with an example (which may also illustrate how Lagrange and Abel worked).

In S_4, let H be the subset

$$\{I, \qquad (1\ 2), \qquad (1\ 3), \qquad (2\ 3), \qquad (1\ 2\ 3), \qquad (1\ 3\ 2)\}.$$

We could check that it is closed under multiplication, but why bother? Those are the only possible permutations in S_4 that do not move 4. It is obvious that H is a copy of S_3, in which we know the multiplication to be closed. Hence H is a subgroup.

Take any permutation missing from H, like (1 4). Denote by $(1\ 4)H$ the subset of S_4 consisting of products $(1\ 4)h$, using an h from H. That subset is called a **left coset** of H. Check, or take my unreliable word for it, that $(1\ 4)H$ consists of

$$(1\ 4), \qquad (1\ 4\ 2), \qquad (1\ 4\ 3), \qquad (1\ 4)(2\ 3), \qquad (1\ 4\ 2\ 3), \qquad (1\ 4\ 3\ 2).$$

(Is that a subgroup?) It has six elements, same as H. With all those 4's, it obviously shares no elements with H. Neither of those facts is an accident.

Each member h of H produces $(1\ 4)h$. No two unequal members can produce equal products: If $(1\ 4)h_1$ equals $(1\ 4)h_2$, then cancellation gives us $h_1 = h_2$. That is why $(1\ 4)H$ has the same number of elements as H. Separately, no member h_1 of H could match a member $(1\ 4)h_2$ of $(1\ 4)H$:

$$h_1 = (1\ 4)h_2 \qquad \text{would force} \qquad h_1h_2^{-1} = (1\ 4)h_2h_2^{-1} = (1\ 4);$$

the left side is in H, and (1 4) is not. That is why H and $(1\ 4)H$ have no elements in common. Therefore H and $(1\ 4)H$ together account for 2×(order of H) members of S_4.

Is anything still unaccounted for? Clearly (2 4) is nowhere so far. The coset (2 4)H comprises

(2 4), (1 2 4), (2 4)(1 3), (2 4 3), (1 2 4 3), (1 3 2 4).

As before, this new coset has (order of H) elements, none in common with H. It is also disjoint from the first coset. If an element (1 4)h_1 of (1 4)H were (2 4)h_2 from (2 4)H, then we would have

$$(1\ 4)h_1 = (2\ 4)h_2, \qquad \text{forcing} \qquad (1\ 4)(h_1h_2^{-1}) = (2\ 4).$$

That would put (2 4) in (1 4)H, contrary to why we chose (2 4).

Without guessing at how many more left cosets there are (How many are there?), we can be sure of one thing. At some point, we will run out of new permutations to create new cosets. At that point, H and its other cosets (H is itself the coset IH) will together fill G. Since the cosets are disjoint subsets with (order of H) members each, the population of G is given by

(order of G) = (order of H)(number of left cosets).

That is the reason for Lagrange's theorem.

Look at the set of nonzero residues modulo some prime. Take the prime 43, so that the set is

$R_{43} = \{1, 2, \ldots, 42\}$.

Apply the operation of multiplication modulo 43. If i and j are in R_{43}, then they are not divisible by 43. Hence neither is their product; ij modulo 43 is another nonzero residue. That makes multiplication an operation in R_{43}. It is associative, and we know 1 is the identity. We learned (section VIII.C.2b(ii)) that every member of R_{43} has a multiplicative inverse mod 43. Therefore R_{43} is a group under multiplication. In it, each member (say 15) has some order k. That k is the order of the cyclic subgroup generated by 15. By Lagrange's theorem, k divides (order of R_{43}) = 42. We draw our favorite conclusion,

$$15^{42} = (15^k)^{42/k} \equiv 1 \mod 43. \qquad \text{(See Exercise 7e.)}$$

Exercises IX.A.2a(iii)

1. In S_4, write all the distinct powers of the given cycle, and verify that the next-to-last is the given one's inverse:
   a) (1 2) b) (1 2 3) c) (1 2 3 4).
2. Show that if g has order k, then for $1 \le j \le k$,
   $(g^j)^{-1} = g^{k-j}$.
   (Is the restriction on j needed?)
3. In our previous example **Z**$_6$, show that {1, 5} is closed under (natural number) multiplication. Is it a subgroup of **Z**$_6$?
4. Is it possible for a group G to have *only* the automatic subgroups G and $\{I\}$?
5. What are the subgroups of S_3? (Hint: There are six, total.)
6. In the group **Z** of integers under addition:
   a) Show that Theorem 3 does not apply.
   b) Describe all the subgroups.
7. Unite number theory and group theory to prove Euler's theorem about the totient function:
   a) List the $\varphi(20) = 8$ numbers below 20 that are relatively prime to 20.
   b) Show using number theory—not calculation—that they form a group under the operation of multiplication modulo 20.
   c) Show using group theory that for each k among them,
   $k^{\varphi(20)} \equiv 1$ modulo 20.
   d) Is this group cyclic?
   e) Is R_{43} cyclic? (Hint: Work from $3^4 \equiv -5 \mod 43$.)

b) groups of permutations

(i) the inclusion

We made an offhand remark that everything about our groups is found in the symmetric groups. We will make it less offhand by showing that every finite group has a twin within a symmetric group.

Theorem 1. If G has order n, then there is a subgroup of order n in S_n in which the multiplication works the same way as the operation in G.

[I am avoiding the technical description: There is within S_n a subgroup "isomorphic" (Greek root for "same form") to G.]

Let's see what that means in our two examples $\mathbf{Z}_6$ and S_3.

We noted that $\mathbf{Z}_6 = \{1, 2, 3, 4, 5, 0\}$ is the cyclic group generated by 1 under addition modulo 6. Certainly S_6 is not cyclic. (Evidence?) However, take within S_6 the cyclic subgroup C generated by the cycle $c = (1\ 2\ 3\ 4\ 5\ 6)$. We saw (Exercise IX.A.1d(iii):1) that the order of a 6-cycle is six. Therefore the correspondence

$$1 \leftrightarrow c, \qquad 2 \leftrightarrow c^2, \qquad \ldots, \qquad 5 \leftrightarrow c^5, \qquad 0 \leftrightarrow c^6 = I$$

matches the six elements of $\mathbf{Z}_6$ with those of C. More important, the multiplication in C reflects the addition in $\mathbf{Z}_6$:

$$i + j \quad \text{modulo } 6 \qquad \leftrightarrow \qquad c^{(i+j \quad \text{modulo } 6)} \qquad = \qquad c^i\, c^j.$$

In the case of S_3, we know its members are

$$I, \qquad (1\ 2), \qquad (1\ 3), \qquad (2\ 3), \qquad (1\ 2\ 3), \qquad (1\ 3\ 2).$$

Match those with the members of S_6

$$I, \qquad (1\ 2)(3)(4)(5)(6), \qquad \ldots \quad , \qquad (1\ 3\ 2)(4)(5)(6).$$

Clearly the S_3 multiplication of its six cycles gives precisely the results, in abbreviated form, of the corresponding products in S_6.

To see why the theorem is true, label the elements of G:

$$G = \{g_1, g_2, \ldots, g_n\}.$$

Let g be one of them. Define the function f_g by

$$f_g(g_i) \;=\; g_i\, g \qquad \text{(the multiplication in } G) \qquad \text{for each } i.$$

The function is a permutation. That is, it is one-to-one: The only way

$$f_g(g_i) \;=\; f_g(g_j) \qquad \text{is} \qquad g_i\, g \;=\; g_j\, g.$$

By cancellation, that forces $g_i = g_j$. The function permutes the n elements of G, is therefore in S_n.

If $h \neq g$ is another element of G, then $f_h \neq f_g$. That they are different permutations is immediate:

$$f_g(I) = g \qquad \text{and} \qquad f_h(I) = h \qquad (I \text{ the identity in } G)$$

are unequal values. That makes the association $g \leftrightarrow f_g$ a one-to-one correspondence.

Finally, multiplication of these permutations in S_n works the way multiplication in G does. Fix two members g and h of G. For every g_i in G,

$$\begin{aligned}
[f_g\, f_h\,](g_i) &= f_h\,(\,f_g\,(g_i)) && \text{(by the definition of permutation multiplication)}\\
&= (g_i\, g)h && \text{(by the definitions of } f_h \text{ and } f_g)\\
&= g_i\,(gh) && \text{(You decide.)}\\
&= f_{(gh)}(g_i).
\end{aligned}$$

That says $f_g\, f_h$ and $f_{(gh)}$ are the same permutation. In that way, the multiplication in the subgroup

$$\{f_{g_1}, \qquad f_{g_2}, \qquad \ldots, \qquad f_{g_n}\}$$

of S_n reflects the multiplication in G. (Why is that set definitely a *subgroup* of S_n?)

(ii) the alternating group and normal subgroups

Suppose f and g are two even permutations in S_n. Each is factorable into an even number of transpositions. Clearly then fg is also an even permutation. We infer that the subset A_n of even permutations is closed under multiplication. By Theorem 3 in (a(iii)), A_n is a subgroup of S_n. It is called the **alternating group** on n elements.

Beginning with $n = 2$, A_n always has half the $n!$ permutations in S_n.

> Suppose we multiply every member of S_n by (1 2). The multiplication turns every even permutation into an odd one, and vice-versa. As we saw just above, such a multiplication gives a one-to-one transformation: The cancellation law implies that if f and g are different, then $f \times (1\ 2)$ cannot be equal to $g \times (1\ 2)$. [Yes, the transformation is a permutation. What symmetric group does it belong to?] It follows that there are exactly as many evens as odds, and A_n has half the permutations.

Implicit in the previous paragraph is that the set $A_n(1\ 2)$ of products

(even permutation)(1 2)

is the set of odd permutations. By analogy with "left coset" (from Lagrange's theorem in subsection 2a(iii)), we call $A_n(1\ 2)$ a **right coset** of A_n. Clearly A_n and $A_n(1\ 2)$ add up to all of S_n; those are the only right cosets of A_n. Similarly, the left coset $(1\ 2)A_n$ is the set of odd permutations; it and A_n are the only left cosets of A_n. Since $A_n(1\ 2)$ and $(1\ 2)A_n$ are the same set, each right coset of A_n is also a left coset.

> In S_3, A_3 consists of
>
> I, (1 2 3), (1 3 2).
>
> Check that the products of those three with (1 2) (the latter on the right) are
>
> (1 2), (2 3), (1 3),
>
> and their products with (1 2) (on the left) are
>
> (1 2), (1 3), (2 3).
>
> The right coset $A_3(1\ 2)$ is the same set as the left coset $(1\ 2)A_3$.
>
> By contrast, the left and right cosets of
>
> $J = \{I, (13)\}$
>
> (Is that really a subgroup of S_3?) do not coincide. For example,
>
> $J(1\ 2) = \{(1\ 2), (1\ 3\ 2)\}$, whereas $(1\ 2)J = \{(1\ 2), (1\ 2\ 3)\}$.
>
> Keep in mind that the left coset $(1\ 2)J$ is the only one the right coset $J(1\ 2)$ could possibly match. They are the only ones that have (1 2).

A subgroup for which every right coset is a left coset is special: We call it a **normal** subgroup.

There are three situations in which it is automatic that subgroup H is normal in group G:

1. H has half the elements in G. In that case, just as in the argument above for A_n, the lone right and left coset other than H is the rest of G.
2. H is either $\{I\}$ or G. In the usual mathematical usage, those are called the **trivial** normal subgroups.
3. G is abelian. In that case, the cosets gH and Hg match element by element.

There is a test for normality, often given as the definition of "normal," that is generally more convenient than matching up the cosets. (Try it in Exercises 1 and 2.)

Theorem 2. The subgroup H is normal in G iff for every member h of H and g in G (irrespective of whether g belongs to H), the product $g^{-1}hg$ is a member of H.

> To see why, observe that $Hg = gH$ means that each element h_1g of the right coset equals some element gh_2 of the left coset. From
>
> $h_1g = gh_2$, we infer $g^{-1}h_1g = g^{-1}gh_2 = h_2$.
>
> The product $g^{-1}h_1g$ is an element of H. The argument is reversible to prove the converse.

> To see it with our subgroup
> $J = \{I, (1\ 3)\}$
> in S_3, simply note that
> $(1\ 2)^{-1}(1\ 3)(1\ 2) = (1\ 2)(1\ 3)(1\ 2) = (2\ 3)$
> is not in J. This is not a normal subgroup.

Exercises IX.A.2b(ii)

1. Show that $\{I, (1\ 2)\}$ and $\{I, (2\ 3)\}$ are not normal in S_3. [Math has not adopted "abnormal."]
2. In S_4, let
   $f = (1\ 2)(3\ 4)$, $g = (13)(2\ 4)$.
   a) Show that
   $f^2 = g^2 = I$ and $fg = gf$.
   b) Argue why $H = \{I, f, g, fg\}$ is a subgroup of S_4.
   c) Argue as follows that H is a normal subgroup:
   Take a transposition of your choice in S_4, and call it t. Show that H has each of
   $t^{-1}ft$, $t^{-1}gt$, $t^{-1}(fg)t$.
   That says H passes the normality test with t. You could test with the other transpositions, but the calculations are similar, owing to symmetry: H has all three possible products of disjoint transpositions. It then follows that H passes the test with any product of transpositions. That makes H a normal subgroup of S_4.)
3. Show that if an abelian group has composite order, then it has some nontrivial normal subgroup (normal subgroup other than $\{I\}$ and the group itself).

(iii) group structure and solution formulas

There is an architecture to groups, a way they are put together, that is related to the solvability of equations. The relation is the reason for this long discussion of groups, and in particular for introducing normal subgroups. It is what Ruffini and Abel established.

Proposition. (The Abel-Ruffini Theorem) The general polynomial equation of degree n is solvable—there is a formula that produces its solutions in terms of roots and arithmetic operations on the coefficients—if and only if the symmetric group S_n of its complex roots is a solvable group.

A "solvable group" is one in which normal subgroups line up a certain way. In the remainder of this subsection, we will define the term and indicate how Galois proved that if $n \geq 5$, then S_n is not solvable. That last, in view of the Abel-Ruffini theorem, proves that there are no solution formulas for the general polynomials of degrees 5, 6, ….

In S_n, A_n is a *maximal* normal subgroup. That is, you cannot squeeze another normal subgroup between them. Indeed, you could not squeeze *any* subgroup, normal or not, between them: The order of that subgroup would have to divide (order of S_n) = $n!$, and (order of A_n) = $n!/2$ is already the biggest possible divisor.

Now, A_n has normal subgroups. (Name one.) Here we mean subgroups that are normal in A_n, not necessarily normal in S_n. (Exercise 1b asks for evidence that you can have J normal in H, with H normal in G, without J being normal in G.) Let J_1 be any maximal one. If J_1 is not simply $\{I\}$, then among the normal subgroups of J_1, there must be a maximal one J_2. If J_2 is not simply $\{I\}$, …. This descent has to come to an end at $\{I\}$. It leaves us with a sequence

$$\{I\} \quad \text{max normal in} \quad J_k \quad \text{max normal in} \quad J_{k-1} \quad \text{max normal in} \quad \ldots \quad J_1 \quad \text{max normal in} \quad A_n \quad \text{max normal in} \quad S_n.$$

Such a sequence is called a **composition series** for S_n.

In S_3, look at the sequence

$\{I\}$, A_3, S_3.

We know A_3 is maximal in S_3. In A_3, $\{I\}$ is the only subgroup (Reason?), and it is normal. Therefore the sequence is a composition series.

In $\mathbf{Z}_6$, we can say the same things about

$\{0\}$, $\{2, 4, 0\}$, $\mathbf{Z}_6$.

Therefore this sequence is a composition series for $\mathbf{Z}_6$.

Those two composition series are pretty much the same, even though S_3 and $\mathbf{Z}_6$ do not have the same, well, composition. There is, however, a series-related difference between the two groups. In S_3, the series above is the only possible one, because there A_3 is the only maximal normal subgroup. Check (Exercise 2) that

$\{0\}$, $\{3, 0\}$, $\mathbf{Z}_6$

is *another* composition series for $\mathbf{Z}_6$. This one is structurally different from the previous, because in this one, the sequence of orders is 1-2-6; in the previous series, the orders were 1-3-6. Generally, composition series are not unique, although different series for one group do have something in common (as in Exercise 3).

One way you can be sure that the normal subgroup H of G is maximal is if the natural number

(order of G)/(order of H)

is prime. That ratio is useful enough to deserve a name: We call it the **index** of H in G. Recall from the Lagrange argument (section IX.A.2a(iii)) that it is the number of either left or right cosets of H.

Exercise IX.A.2b(ii):2 had the example

$H = \{I,$ $(1\ 2)(3\ 4)$, $(1\ 3)(2\ 4)$, $(1\ 4)(2\ 3)\}$.

It is a normal subgroup in S_4, and therefore normal in A_4. (See Exercise 1a. Why is it even a *subgroup* of A_4?) In A_4, it has prime index 12/4 = 3.

Suppose J is a subgroup of A_4 that contains H. Then the index of H satisfies

3 = (order of A_4)/(order of J) × (order of J)/(order of H).

By Lagrange's theorem, those ratios are integers. Of necessity, one of them is 1. If the first is 1, then $J = A_4$. If it is the second, then $J = H$. That shows H is maximal in A_4.

If a group has a composition series in which each subgroup has prime index in the next one, then we say the group is **solvable**. That is the property Ruffini and Abel connected to solution formulas.

In $S_2 = \{I, (1\ 2)\}$, we have the lone composition series

$\{I\}$, S_2.

It carries the single index 2. For S_3, we saw the composition series

$\{I\}$, A_3, S_3,

with indices 3, 2. Both groups are solvable, and we know formulas exist for quadratics and cubics.

In S_4, we know A_4 is normal, with index 2. We know

$H = \{I,$ $(1\ 2)(3\ 4)$, $(1\ 3)(2\ 4)$, $(1\ 4)(2\ 3)\}$

is normal in A_4, with index 3. We see that

$J = \{I,$ $(1\ 2)(3\ 4)\}$

is a subgroup of H (Why?), has index 2, is therefore normal and maximal in H. Last, $\{I\}$ has index 2 in J. We have the composition series

$\{I\}$, J, H, A_4, S_4,

with all four indices prime. Therefore S_4 is solvable.

Evidently, we are going to find that S_5 is not solvable. Ruffini and Abel proved it by working directly with the permutations. Galois, in the material Liouville published in 1846, worked more with group properties. His proof is based on the next proposition.

Proposition. For $n \geq 5$, the alternating group A_n is **simple**: Its only normal subgroups are the automatic two, A_n and $\{I\}$.

The proof is actually elementary, just more detailed than we want to tackle. You can find one at Georgia Tech. It hangs on two facts: Every normal subgroup of A_n ($n \geq 3$) includes a tricycle; and if a normal subgroup of A_n has one tricycle, then it has them all. Because every even permutation factors into the product of tricycles (Exercise 4), a subgroup with all of them has to be all of A_n.

We have met simple groups before. In answer to Exercise IX.A.2a(iii):5, if a group has prime order, then its only subgroups are $\{I\}$ and the group itself. Hence a group of prime order is necessarily simple. As it happens, those are the only (finite) *abelian* groups that are simple (Exercise 3 in subsection (ii)).

Accepting the proposition, we can prove the following.

Theorem 3. For $n \geq 5$, the only nontrivial normal subgroup of S_n is A_n.

> Given $n \geq 5$, let J be a normal subgroup of S_n. Then the intersection $J \cap A_n$ is a normal subgroup of S_n, is therefore a normal subgroup of A_n (both conclusions from Exercise 5). Because A_n is simple, $J \cap A_n$ has to be either all of A_n or $\{I\}$.
>
> Suppose first $J \cap A_n = A_n$. That means A_n is a *subset* of J, therefore a subgroup of it. We know that you cannot squeeze a subgroup between A_n and S_n. It must be that either $J = A_n$ or $J = S_n$.
>
> Suppose instead that $J \cap A_n = \{I\}$. That says J has no even permutations except I. It will then be impossible for J to have any odd ones. Imagine f is an odd permutation n J. Then the even permutation f^2 is in H, forcing $f^2 = I$; f has order 2. By Exercise IX.A.1d(iii):1, the factorization of f into disjoint cycles has only disjoint *transpositions*. Say
>
> $f = (1\ 2)$ (even number of disjoint transpositions devoid of 1 and 2).
>
> Because J is normal, it must also have the odd permutation
>
> $g = (1\ 3)^{-1} f\,(1\ 3) = (3\ 1)\,(1\ 2)$ (disjoint transpositions devoid of 1 and 2) $(1\ 3)$.
>
> This g is different from f. It is different because f assigns
>
> 2 → 1 → (stays 1 through the red transpositions),
>
> whereas g assigns
>
> 2 → 1 → (stays 1 through the red transpositions) → 3.
>
> Because g is not $f = f^{-1}$, fg is another non-identity element of J, and fg is *even*. By that contradiction, we conclude J has no odd permutations. We infer $J = \{I\}$. That demonstrates the theorem.

Now that we know all the normal subgroups of S_n for $n \geq 5$, we can show that S_n has only one composition series. The only maximal normal subgroup of S_n is A_n, and by the proposition, the only normal subgroup of A_n is $\{I\}$. Therefore the only possible composition series for S_n is

$$\{I\}, \quad A_n, \quad S_n.$$

In it, the first index is

$$n!/2 = [1(2)\ldots5(6)\ldots n]/2 = 60 \times \text{something}.$$

That number is not prime. Hence S_n is not solvable. There is no solution formula for the general polynomial equation of degree n.

Like Ruffini and Abel, Galois answered a question from the old algebra, the theory of equations of all those great names from al-Khwarizmi through Viète. But Galois answered it by creating the new algebra, the theory of abstract algebraic structures.

[There is a must-read book I recommended at the opening of Section VI.B. Galois and Abel are at its center. It is Mario Livio's *The Equation That Couldn't Be Solved*.]

Exercises IX.A.2b(iii)

1. a) Show that if J is a subgroup of H, and both are normal subgroups of G, then J is a normal subgroup of H.
   b) Give an example in which J is a normal subgroup of H, and H is a normal subgroup of G, but J is not normal in G. (Hint: An example exists in this subsection.)
2. Show that
   $\{0\}$, $\{3, 0\}$, $\mathbf{Z}_6$
   is a composition series for $\mathbf{Z}_6$. (Note: There are several things to prove.)
3. We [you] saw in Exercise IX.A.2a(iii):7e that the group R_{43}, of residues modulo 43 under multiplication, is cyclic.
   a) How many composition series does it have?
   b) For each series, what is the set of *indices* (of each subgroup in the next one)?
4. Show that every even permutation is the product of tricycles. (Hint: The product of *two* transpositions has to be I, or a tricycle, or the product of two tricycles.)
5. Show that if H and J are normal subgroups of G, then:
   a) $H \cap J$ is a normal subgroup of G. (Note: There are two things to prove.)
   b) $H \cap J$ is a normal subgroup of both H and J.

c) Galois theory and the ancient problems

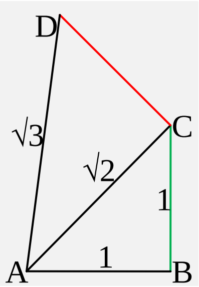
Galois theory encompasses a collection of beautiful results. We will focus on just one, because it leads to the end of a two-millennium geometric hunt. It is a technical mouthful, so we will state it, then chew it in small bites.

Proposition. A number is constructible if and only if it is algebraic and has minimal polynomial whose degree is a power of 2.

(i) constructible numbers

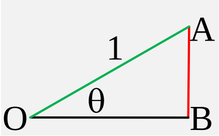
Recall from section III.A.3 the Greek challenge to construct figures with various properties. By **constructing a number**, we mean producing a line segment of that length by resort to three weapons: a compass, a straightedge, and a unit length.

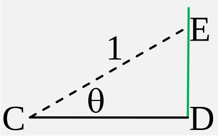
Think of a straightedge as an unmarked ruler. You cannot measure with it; you cannot produce, for example, a one-inch segment. However, informed that this underline is one inch long, you can use the compass and straightedge to reproduce the segment. Then you can adjoin consecutive congruent segments to draw a segment of any natural-number length. Therefore natural numbers are constructible. We know there is a construction to partition a length *m* into *n* congruent pieces. Therefore positive rational numbers are constructible. If we agree to consider lengths in one sense along a line to be positive, in the opposite sense to be negative—and a point to have zero length—then we see that all rational numbers fit the bill. They must be among the numbers the proposition identifies.



At least some irrational numbers are constructible. At right, we have AB of unit length. We raise the perpendicular (green) at B, and make BC also 1 long. Then the length of AC is √2. Erecting the perpendicular CD (red) of unit length to AC, we have AD of length √3. Continuing, we construct every √*n*.

The reason for studying constructible numbers is their relation to the three ancient problems. We can construct a unit circle. To square it is to produce a square of area π. That amounts to constructing a side of length $\sqrt{\pi}$. Similarly, doubling a unit cube amounts to constructing a side of length $\sqrt[3]{2}$. "Trisecting an angle" refers to producing an angle, but we can turn it into constructing a length. Suppose the angle to divide has measure 3θ. Trisecting it, to make an acute angle θ, is equivalent to constructing either sin θ or cos θ. If we are given θ with vertex O, as at left, we lay off length 1 (green) to A along one side, then drop the perpendicular (red) to B on the other. That constructs sin θ = AB and cos θ = OB. If conversely we are given say cos θ, the length CD at right, then we build the perpendicular (green) at D, then use the compass to mark the point E on the green where CE = 1. That constructs angle ECD, whose cosine is cos θ.





(ii) fields

We need to define one more abstract algebraic structure. A set *K* is called a **field** when it is equipped with *two* operations that satisfy eleven axioms. The axioms correspond to eleven properties of the field of rational numbers, or of the field of real numbers. Accordingly, the operations are called **addition** and **multiplication**, denoted by + and × (the latter sign sometimes left out). The rules are:

Axioms 1-5. Under addition, *K* is an abelian group.

The "1-5" is a way to indicate that addition satisfies the four group axioms, plus the requirement of commutativity. Thus, + is an operation; it is associative; it has an identity, denoted by 0 and called "zero"; and every element *a* of *K* has an additive inverse, denoted by -*a* (and hard to name. "Negative *a*" has the disadvantage that we need "negative" for a later use. Similarly, "*a* inverse" is better saved for multiplication. "The additive inverse of *a*" has too many notes. "Minus *a*" will do, although "*a* opposite" is apposite.) The fifth requirement is $a + b = b + a$.

[Richard Dedekind, whom we will meet later, described fields and called them by the German word for "body," *Körper*. Perhaps because *F* is too valuable in connection with functions, math has kept *K* as the symbol for a field. *G* is valuable in the same context, but the French word for "group" is cognate to the English word; "group *G*" stays.]

Axioms 6-10. Under multiplication, the nonzero elements of *K* form an abelian group.

Although multiplication by 0 has to be defined, Axioms 6-10 say nothing about it. They demand something of the other elements. There *must be* other elements, because they form a group, and a group cannot be empty. Among the nonzero elements, × has to be an operation; it must be associative; there must exist a multiplicative identity, designated by 1 and called "one"; each element *b* must have a multiplicative inverse, written b^{-1} and called "*b* inverse"; and $ab = ba$.

The two sets 1-5 and 6-10 of axioms are almost symmetric. They differ only in the odd exemption given, with respect to multiplication, to 0. More important, nothing so far connects addition and multiplication. The last rule does that.

Axiom 11. Multiplication is distributive over addition: For any elements *a*, *b*, *c* of *K*,

$$a(b + c) = ab + ac.$$

In that last equation, we have adopted the usual convention about order of operations.

From those axioms, what can we prove?

Theorem 1. For any element *a* in any field,

$$a \times 0 = 0 \times a = 0.$$

We have

$$\begin{aligned} a \times 0 + 0 &= a \times 0 && \text{(identity property of zero)} \\ &= a \times (0 + 0) && \text{(same)} \\ &= a \times 0 + a \times 0 && \text{(distributivity).} \end{aligned}$$

In every group, cancellation applies. Cancel $a \times 0$ to arrive at

$$0 = a \times 0.$$

For $0 \times a$, just mirror-image the argument.

Theorem 2. In any field, 0 *cannot have a multiplicative inverse*, and addition cannot be distributive over multiplication (Exercises 1 and 2).

Theorem 3. In any field, if $ab = 0$, then either $a = 0$ or $b = 0$.

Theorem 1 says that if one of the factors is zero, then the product is zero. Theorem 3 says the converse. That is a fundamental algebraic principle. In old (equations) algebra, we solve equations by factoring expressions we know to be zero and reducing the question to what makes the factors zero.

Assume $ab = 0$. If $a \neq 0$, then a has an inverse a^{-1}. It follows that

$$\begin{aligned} 0 &= a^{-1} \times 0 && \text{(Theorem 1)} \\ &= a^{-1} \times (ab) && \text{(by assumption)} \\ &= (a^{-1} \times a)b && \text{(associativity)} \\ &= 1b = b && \text{(definitions of inverse and identity).} \end{aligned}$$

If $ab = 0$, either a is 0, or b has to be.

Almost all of our operations-based algebra—the pre-college “algebra” of our schools—follows from the axioms. On one side, take manipulations. The dreaded rules for adding and dividing fractions,

$$a/b + c/d = (ad + bc)/(cd), \qquad (a/b)/(c/d) = (ad)/(bc),$$

are easy to prove from the axioms (after some needed definitions; see Exercise 3a-c). On the other side, take equations and solutions methods. The “general linear equation”

$$ax + b = 0 \qquad \text{with } a \neq 0$$

always has exactly one solution, given by $x = -b/a$ (Exercise 3d). For another example, take the quadratic

$$ax^2 + bx + c = 0, \qquad a \neq 0.$$

If $b^2 - (1 + 1)^2ac$ has a **square root**, an element d such that $d^2 = b^2 - (1 + 1)^2ac$, then

$$x = (-b \pm d)/(a + a)$$

gives the *only* two (possibly equal) solutions. (The proof has many steps—for example, you have to establish that multiplying binomials works the way we expect—but they are as elementary as they are in our usual proof of the quadratic formula. See Exercise 4 about the existence of square roots.)

We said more than once “in any field.” There *are* fields other than those of rational, real, or complex numbers. Look at $\mathbf{Z}_{43}$, the set $\{0, 1, 2, \ldots, 42\}$ of residues modulo 43 with addition and multiplication modulo 43. Under that addition, $\mathbf{Z}_{43}$ is an abelian group; the explanation matches that for $\mathbf{Z}_6$ in section IX.A.2a(ii). The nonzero residues form the abelian group R_{43} (end of section IX.A.2a(iii)) under multiplication. We trust that the multiplication is distributive, for a reason we have seen before: Multiplication of integers is distributive, and congruence is compatible with it and addition. (It should be clear that all these statements hold in $\mathbf{Z}_p$, for any prime p.)

Those fields $\mathbf{Z}_{43}$, $\mathbf{Q}$ (rationals), $\mathbf{R}$ (reals), $\mathbf{C}$ (complex) really are different algebraically. In $\mathbf{Z}_{43}$ but not the other three,

$$1, \qquad 1 + 1, \qquad 1 + 1 + 1, \qquad \ldots$$

lists finitely-many elements. (We have seen why.) In $\mathbf{Q}$ but not the other three,

$$x^2 - 6 = 0 \qquad \text{(“6” meaning } 1 + 1 + 1 + 1 + 1 + 1)$$

> has no solution. (What is the solution in $\mathbf{Z}_{43}$?) In **R** and not the others,
>
> $$x^3 - 21 = 0$$
>
> has exactly one solution? (What are the solutions in $\mathbf{Z}_{43}$ [recall that R_{43} is cyclic] and **C**?) Finally, in **C** and not the others,
>
> $$x^2 + 1 = 0$$
>
> has a solution. (The $\mathbf{Z}_{43}$ case is about square roots, Exercise 4d.)

Exercises IX.A.2c(ii)

1. Prove that in a field, 0 cannot have a multiplicative inverse. (The axioms do not require 0 to have an inverse. Here we say that they actually prohibit an inverse.)

2. Prove that in a field, addition cannot distribute over multiplication:

   $$a + (b \times c) = (a + b) \times (a + c)$$

   cannot always (for all elements of the field) be true.

3. a) Based on the axioms, define $a - b$ and $a \div b$ (hereafter a/b).
   b) Prove (in a field) the familiar algebraic relation

   $$(a - b)(a + b) = a^2 - b^2.$$

   c) Show that

   $$(a/b)/(c/d) = (ad)/(bc).$$

   State the needed conditions.
   d) Show that $x = -b/a$ solves

   $$ax + b = 0,$$

   and nothing else does.

4. a) Show that in any field, 1 has at least one square root.
   b) In (a), is it possible for there to be *just* one square root? Is it possible for there to be three (distinct) square roots?
   c) Give one example of a finite field in which -1 has no square root; a second example in which it has exactly one; a third in which it has two. Is more than two possible?
   d) In $\mathbf{Z}_{43}$, does -1 have a square root? (Hint: R_{43} is *still* cyclic with 42 elements, and 42 is not divisible by 4.)

(iii) fields of real numbers

Within the field **R** of real numbers, the subset **Q** of rational numbers is itself a field under the addition and multiplication in **R**. We therefore say **Q** is a **subfield** of **R**.

By analogy with our study of maximal subgroups (section IX.A.2b(iii)), we may ask whether there are subfields smaller than **Q**, or between **Q** and **R**. The answers are no and yes.

> Suppose K is a subfield of **R**. Then K has to possess an additive identity O. That identity must have
>
> $$O + O = O$$
>
> under the addition in K, which is the addition in **R**. Since the latter addition allows cancellation, we conclude that $O = 0$. Additionally, K must have a multiplicative identity I that must be different from $O = 0$ and must satisfy
>
> $$I \times I = I.$$
>
> Because I is nonzero, we may cancel it to conclude $I = 1$. We have so far 0 and 1 in K.
>
> We must also have 1 + 1, 1 + 1 + 1, … in K. That says the natural numbers are in K, from which their additive inverses are in K. Thus, all integers are in K. Hence all rational numbers mn^{-1} are also members. We deduce that if K is a subfield of **R**, then K contains **Q**; no subfield is smaller than **Q**.

The larger-field question is what Dedekind was actually studying. Pick a convenient real number outside **Q**, like √2. Let *S* be the subset of real numbers of the form *r* + *s*√2, where *r* and *s* are rational. This subset is closed under subtraction and division: If *a* = *r* + *s*√2 and *b* = *t* + *u*√2 ≠ 0 are in *S*, then so are *a* – *b* and *a*/*b* (Exercise 1a). We will show below that those properties guarantee *S* is a subfield of **R**. As such, it is a bigger subfield than **Q** and smaller than **R**. (See Exercise 1b-c, as well as Exercise 2.)

[Let us call the members of *S* "the surd numbers." "Absurd" would be a decent substitute. Normally the word is applied to any irrational combination of roots. It comes from the Latin for "deaf," a usage that appears to go all the way back to al-Khwarizmi. He called such numbers "silent."]

Recall the theorem (section IX.A.2a(i)) that if a subset of a finite group is closed under the operation, then the subset is a subgroup. The set

{1, 1 + 1, 1 + 1 + 1, …}

of natural numbers is closed under both operations in **R**, but is not a subfield (Exercise 3). The subfield criterion corresponding to that group theorem is the following.

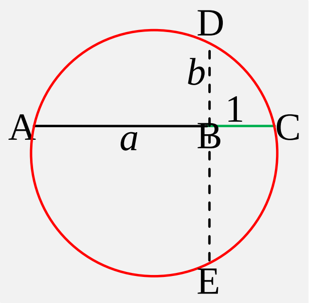
Theorem 4. If a subset of a field has at least one nonzero element and is closed under subtraction and division, then it is a subfield.

> Assume *T* is such a subset and *a* is any nonzero element in *T*. Then 0 = *a* – *a* and 1 = *a*/*a* are in *T*, satisfying the identity axioms. Therefore, 0 – *a* = -*a* and 1/*a* = a^{-1} are also in *T*, satisfying the inverse axioms. If *b* ∈ *T* (zero or not), then *b* – -*a* = *a* + *b* and *b*/a^{-1} = *ab* are in *T*, so that addition and multiplication are (closed) operations in *T*. Associativity, commutativity, and distributivity are inherited from the parent. Thus, *T* satisfies the field axioms under the overlying operations.

In this context, our primary interest is that the constructible numbers form a subfield. Assume that we can construct *a* and *b*. We easily construct *a* – *b*, with our previous convention about positive lengths toward one side and negative toward the other. For division, extend segment AB of length *a* by 1 (green) to C, as at right. Draw the perpendicular (dashed) at B, and match length *b* from B to D. Construct the circumscribed circle for triangle ACD, intersecting the other side of the perpendicular at E. Then the intersecting-chords property says

a(1) = *b*(BE).

We have constructed BE = *a*/*b*. By Theorem 4, it follows that the constructible numbers constitute a subfield of the reals. (See also Exercise 4.)



Exercises IX.A.2c(iii)

1. a) Show that if *a* = *r* + *s*√2 and *b* = *t* + *u*√2 ≠ 0 are surds, then so are *a* – *b* and *a*/*b*.
   b) Show that **Q** is a proper subset (not all of) the field *S* of surds.
   c) Show that *S* is a proper subset of **R**. (Hint: Find a real non-surd.)
2. Let *U* be the subset of real numbers of the form *r* + *s*√2 + *t*√3 + *u*√6, with *r*, *s*, *t*, and *u* all rational. Show that *U* is a subfield of **R** bigger than *S* (the surds). (Hint:
   r + *s*√2 + *t*√3 + *u*√6 = (*r* + *s*√2) + (*t* + *u*√2)√3,
   and we already know the surds make up a subfield.)
3. Show that the set of natural numbers is closed under addition and multiplication in **R**, but is not a subfield of **R**?
4. Given a segment of length *a* > 0, construct √*a*.

(iv) algebraic numbers and squaring the circle

At the beginning of section VIII.D.1, we cited the need to divide coefficients as reason to allow polynomials with rational coefficients. It turns out that the important thing is to restrict coefficients *to a field*; any field will do.

> In the set of "polynomials over a field," we can write the induction proof we suggested back there for the division algorithm. From the division algorithm, we can prove the remainder theorem. (Look back at our proof. Observe that the key was setting $x = u$ in
>
> $f(x) = q(x)(x – u) + v$.
>
> The substitution makes $(x – u) = 0$. Consequently
>
> $f(u) \ = \ q(u)0 + v \ = \ v$,
>
> because multiplication by 0 gives 0 in every field. That established the remainder theorem.) From the remainder theorem, we prove the factor theorem.

A member of a field is said to be **algebraic over a subfield** if it is a root of some polynomial whose coefficients are in the subfield. Our interest is just **R** and **Q**, so we will simply say that a real number **is algebraic** if it **satisfies** (is a root of) some polynomial with rational coefficients.

> Every rational number is algebraic (Exercise 1). Each of the irrationals $\sqrt{2}$, $\sqrt[3]{2}$, and cos 20° ≈ 0.94 is algebraic:
>
> | | | | |
> |---|---|---|---|
> | $\sqrt{2}$ | satisfies | $x^2 – 2$, | |
> | $\sqrt[3]{2}$ | satisfies | $x^3 – 2$, | |
> | cos 20° | satisfies | $8x^3 – 6x – 1$ | (section VI.C.4a). |
>
> (It is ironic that for this definition, we could demand *integer* coefficients. If r solves
>
> $x^{1234}/56 + 78x^{910}/1112 – 1314/1516 \ = \ 0$,
>
> then it solves the same equation multiplied on both sides by 56(1112)1516.)

A number that is not algebraic is called **transcendental**. Proving transcendental numbers even exist is an advanced problem. Liouville proved their existence in 1844, then constructed one in 1851. In 1861, Charles Hermite established that e is transcendental. In 1882, Ferdinand von Lindemann exhibited a collection of transcendentals. (Wikipedia® has a [hopelessly advanced] proof.) One member of that collection was π.

View that fact in light of the Galois (constructible-number) proposition. Given that π is not algebraic, we infer that it is not constructible. Therefore $\sqrt{\pi}$ is not constructible. (Why?) Since $\sqrt{\pi}$ is not constructible, it is impossible to construct a square of area π. Hence you cannot square the unit circle.

In that way, a conclusion from the algebra of abstract structures answered one of the geometric puzzles Anaxagoras proposed some 2300 years before.

Exercises IX.A.2c(iv)

1. a) Prove that every rational number is algebraic.
   b) Show that the Galois proposition guarantees every rational number is constructible.
2. Show that $\sqrt[4]{5}$ is algebraic.
3. Show that $\sqrt[6]{7/8 + \sqrt[9]{10/11}}$ is algebraic.

(v) minimal polynomials and the other problems

An algebraic number will satisfy a whole family of polynomials over **Q**. Because √2 solves

$$x^2 - 2 = 0,$$

it also solves

$$(x^2 - 2)^2 = 0 \qquad \text{and} \qquad (x^2 - 2)(x^8 + 7x^6 - 5x^4 + 3) = 0;$$

it satisfies any multiple of $x^2 - 2$. By the well-ordering principle, among the polynomials √2 satisfies, there must be some of smallest degree. Let $g(x)$ be one of them. We will call $g(x)$ **a minimal polynomial** for √2.

Theorem 5. If g is a minimal polynomial for the real number b, then g divides every other polynomial that b satisfies.

> Assume g is minimal for b, and f is another polynomial b satisfies. By the division algorithm,
>
> $$f(x) = q(x)g(x) + r(x),$$
>
> with $r(x)$ either zero or of smaller degree than $g(x)$. Substitute $x = b$ to rewrite
>
> $$r(b) = f(b) - q(b)g(b) = 0.$$
>
> That says b also satisfies $r(x)$. We conclude that r cannot have a degree: It must be $r = 0$. Therefore g is a divisor of f.

Now suppose $h(x)$ ties $g(x)$ for smallest degree. We still have

$$h(x) = Q(x)g(x).$$

Of necessity, Q is of degree 0; $Q(x)$ has constant value k. We conclude that every minimal polynomial is a nonzero *numerical* multiple of g. Therefore among the minimal polynomials, there is a unique **monic** (leading coefficient = 1) one. We will call that one **the minimal polynomial** for b.

What is the minimal polynomial for √2? If it were not $x^2 - 2$, then by Theorem 5, it would be some smaller-degree factor of it. That factor would be a linear $x - r$ with a rational r. By the factor theorem, that r would solve

$$x^2 - 2 = 0.$$

There is no such rational. Therefore $x^2 - 2$ is the minimal polynomial for √2. (Compare Exercise 1.)

The real number √2 is algebraic, and its minimal polynomial has degree 2. By the constructible-number proposition, √2 is constructible.

> Let $f(x)$ be the minimal polynomial for $\sqrt[3]{2}$. Then by Theorem 5, f divides $x^3 - 2$:
>
> $$x^3 - 2 = q(x)\,f(x).$$
>
> The degrees of f and q have to add up to 3. Neither of them can be the linear $x - r$, because that would imply a rational r whose cube is 2. Therefore one of them has zero degree. Necessarily it is q: $q(x) = 1$, and the minimal polynomial for $\sqrt[3]{2}$ is $x^3 - 2$.
>
> The real number $\sqrt[3]{2}$ is algebraic, with minimal polynomial whose degree is not a power of 2. By the proposition, $\sqrt[3]{2}$ is not constructible. You cannot double the unit cube.

That leaves the problem of trisecting the angle. We know how to construct a 60° angle. As we noted in subsection (i), trisecting it is equivalent to constructing cos 20°. We know cos 20° is one solution of

$$F(x) = 8x^3 - 6x - 1 = 0.$$

By the proposition, if F is minimal for cos 20°, then that number is not constructible. By our previous reasoning (for $\sqrt[3]{2}$), F will be minimal if it has no rational roots. The trisection question has come down to a result of separate interest.

Theorem 6. (The Rational Roots Theorem) Suppose m/n is a lowest-terms rational root of a polynomial with *integer* coefficients. Then m divides the constant term and n divides the leading coefficient.

> To prove it, suppose m/n is reduced and solves the equation
> $$a_k x^k + a_{k-1} x^{k-1} + \dots + a_1 x + a_0 = 0,$$
> in which each a_j is an integer. Substitute $x = m/n$, multiply by n^k, and rewrite the result as
> $$a_k m^k + a_{k-1} m^{k-1} n + \dots + a_1 m n^{k-1} = -a_0 n^k.$$
> The left side is divisible by m. That means m divides $a_0 n^k$. Because m/n is reduced, m is relatively prime to n. Therefore m is relatively prime to all powers of n. Since m divides $a_0 n^k$, we conclude that m divides a_0. If instead we rewrite the last equation as
> $$a_k m^k = -a_{k-1} m^{k-1} n - \dots - a_1 m n^{k-1} - a_0 n^k,$$
> then we reason similarly to conclude that n divides a_k.

Look back at Exercise V.B.2:4. The argument needed there for Fibonacci's cubic is exactly as above, just in a specific setting. Both Leonardo and Omar must have known the idea underlying the rational roots theorem. (See Exercise 3 below.)

Notice that the theorem does not find the solutions of an equation. It simply reduces the candidates to a manageable set. [We would have said "*field* of candidates," but the word is taken.]

> We were chasing
> $$F(x) = 8x^3 - 6x - 1 = 0.$$
> The theorem says that if m/n is a reduced rational solution, then m divides -1 and n divides 8. That forces
> $$m = \pm 1 \qquad \text{and} \qquad n = 1, 2, 4, \text{ or } 8.$$
> (Attaching the sign to m obviates the need to consider negative n.) Therefore the only possible rational solutions are the eight numbers ±1/1, ±1/2, ±1/4, ±1/8. None of those works (Exercise 4). We infer that $F(x)$ does not have rational roots.

That F has no rational roots tells us that it is the minimal polynomial for cos 20°. By the proposition, the real number cos 20° is not constructible. Therefore you cannot construct a 20° angle. You cannot trisect a 60° angle.

Exercises IX.A.2c(v)

1. Show that if b is algebraic, then its minimal polynomial is irreducible (not factorable into polynomials, having rational coefficients, of lower nonzero degree).
2. How did the algebraic problems studied by al-Khwarismi differ from those studied by Galois? How did their methods differ?
3. Use the rational roots theorem to show that Fibonacci's cubic
$$x^3 + 2x^2 + 10x - 20 = 0$$
has no rational solutions.
4. Show that none of ±1/1, ±1/2, ±1/4, ±1/8 satisfies $8x^3 - 6x - 1$.
5. Is $\sqrt[4]{5}$ constructible? Give Galois's existential answer and a constructive answer.
6. Is it possible to construct an angle of:
a) 1° b) 2° c) 3°?
d) On the basis of those, find all the whole-number-degree angles that are constructible.
7. For each number, describe how to construct a regular polygon of that many sides, or argue why it is impossible:
a) 10 b) 18 c) 30 d) 36 e) 48?

Section IX.B. The Calculus

By 1820, the calculus was century-and-a-half old. In that time, nobody had settled d'Alembert and Berkeley's objections to the ambiguous use of infinitesimals (section VIII.B.3a). Even greater ambiguity attached to the treatment of series as sums, despite paradoxes that arose here and there. In this section, we will see how the foundations of calculus—the principles that made it axiomatic—were laid.

1. The Paradoxes

We already noted the contrary values of "infinitesimal" quantities. The expression that became a derivative was the slope

$$[f(a+h)-f(a)]/h.$$

There, h is not allowed to be zero, except that Fermat found it convenient to set it to zero. In the integral problem, Leibniz summed regions of zero width, sort of, to reach a positive area.

For series, some paradoxes are not subtle. Recall (section V.B.3b) that

$$t = 1 - 1 + 1 - 1 + \dots$$

leads to both $t = 1/2$ and $t = 0$. We also have an equation from Euler,

$$\dots + x^{-3} + x^{-2} + x^{-1} + x^{0} + x + x^{2} + x^{3} + \dots \;=\; 0.$$

At least, this one is easy to break (Exercise 1).

The early practitioner with the best feel for series was Jacques Bernoulli. He concluded that the reciprocal-square series

$$1/1^2 \quad + \quad 1/2^2 \quad + \quad 1/3^2 \quad + \quad 1/4^2 + \dots$$

represents a number, even if he could not evaluate it. He reasoned that it is smaller term by term than

$$1 \quad + \quad 1/1(2) \quad + \quad 1/2(3) \quad + \quad 1/3(4) + \dots,$$

which adds up to 2 (based on Exercise VIII.B.3:2). He rediscovered Oresme's argument for the harmonic series (also section V.B.3b), then reasoned by comparison that the bigger series

$$T = 1/\sqrt{1} + 1/\sqrt{2} + 1/\sqrt{3} + \dots$$

must likewise sum to infinity.

> About this last series, he noted a paradox. Separate odds and evens to write
>
> $$\begin{aligned} T &= (1/\sqrt{1} + 1/\sqrt{3} + 1/\sqrt{5} + \dots) \quad + \quad (1/\sqrt{2} + 1/\sqrt{4} + 1/\sqrt{6} + \dots) \\ &= (1/\sqrt{1} + 1/\sqrt{3} + 1/\sqrt{5} + \dots) \quad + \quad (1/\sqrt{2})T. \end{aligned}$$
>
> Then
>
> $$(1/\sqrt{1} + 1/\sqrt{3} + 1/\sqrt{5} + \dots) \quad = \quad (1 - 1/\sqrt{2})T \quad \approx \quad 0.3T.$$
>
> It is not surprising that the odd terms add up to less than the total. However, you would expect them to contribute more than half; each odd term is bigger than the even one right after it.

Paradoxes aside, pretending that series are sums had yielded such important mathematics as the binomial theorem and Euler's complex exponential. The latter was not Euler's strangest result. He evaluated Frère Jacques's reciprocal-square series by analyzing the polynomial

$$p(x) = 1 - x/3! + x^2/5! - x^3/7! + \dots.$$

> Look at the series for sine,
>
> $$\sin t \;=\; t - t^3/3! + t^5/5! - t^7/7! + \dots.$$
>
> Divide by t and set $x = t^2$, to write
>
> $$(\sin \sqrt{x})/\sqrt{x} \;=\; 1 - x/3! + x^2/5! - x^3/7! + \dots.$$

On that left side, $x \le 0$ is not legal. On the right, it is a perfectly valid substitution. If $x \le 0$, then

$$p(x) \quad = \quad 1 \quad + \quad (-x)/3! \; + \quad (-x)^2/5! \quad + \quad (-x)^3/7! + \ldots$$

is a sum of nonnegative terms that are no greater than the terms in

$$e^{-x} \quad = \quad 1 \quad + \quad (-x)/1! \; + \quad (-x)^2/2! \quad + \quad (-x)^3/3! + \ldots.$$

Admittedly, $p(x)$ is long for a polynomial. Nevertheless, if series are sums, then p must have polynomial properties. One of those, Exercise VIII.D.3a:3 , is that the sum of the reciprocals of its roots must be the negative of (linear coefficient divided by constant), irrespective of the degree or leading coefficient. (That last is a good thing, since p has neither.)

For p,

$$-(\text{linear/constant}) \;=\; -[-1/3!]/1.$$

The only possible roots are positive; we noted that if $x \le 0$, then $p(x)$ is 1 + a sum of nonnegative terms. Therefore the roots of p are the positive solutions of

$$(\sin \surd x)/\surd x \;=\; 0.$$

Those are the values of x with

$$\surd x = \pi, 2\pi, 3\pi, \ldots, \qquad \text{namely} \qquad x = \pi^2, (2\pi)^2, (3\pi)^2, \ldots.$$

By the reciprocal-sum property,

$$1/6 \quad = \quad 1/\pi^2 + 1/(2\pi)^2 + 1/(3\pi)^2 + \ldots, \qquad \text{and}$$

$$\pi^2/6 \quad = \quad 1/1^2 + 1/2^2 + 1/3^2 + \ldots.$$

Exercises IX.B.1

1. Use the formula for geometric series to write

$$1/(1 - x) \quad = \quad 1 + x + x^2 + x^3 + \ldots \qquad \text{and}$$

$$[1/x]/(1 - [1/x]) \quad = \quad [1/x]\big(1 + [1/x] + [1/x]^2 + \ldots\big).$$

Add them to write

$$\begin{aligned} 0 \quad &= \quad 1/(1 - x) + 1/(x - 1) \\ &= \quad 1/(1 - x) + [1/x]/(1 - [1/x]) \\ &= \quad (1 + x + x^2 + x^3 + \ldots) + (x^{-1} + x^{-2} + x^{-3} + \ldots). \end{aligned}$$

Now, resolve the paradox.

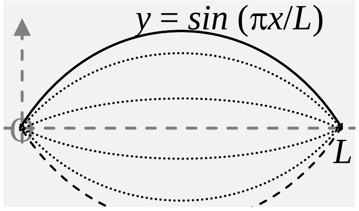
2. Fourier and The Fire

The concerns over imprecise definitions came to a boil when (Jean-Baptiste) Joseph Fourier explained the movement of heat.

Fourier (1768-1830) had an interesting career. He was a scientific advisor under Napoleon in the Egyptian conquest (1798). There, Napoleon named him to head the newly-created Egypt Institute. Back in France, Napoleon rewarded him in 1801with a prefecture Fourier did not want. In was in Grenoble, at the Alps; he would have much preferred to return to his previous position, Lagrange's old post at the *Polytechniqe*. At Grenoble, he met Jean-François Champollion. Read Wikipedia® for an account of how he introduced the Rosetta stone to its eventual decipherer.

a) the background in vibrations

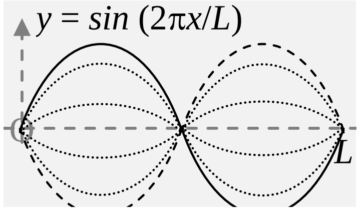
For some perspective, it helps to go back to Daniel Bernoulli. Recall (section VIII.B.1b) that he gave solutions to the string equation, the partial differential equation that governs the possible shapes of a vibrating string. His solutions had been combinations of sines and cosines.

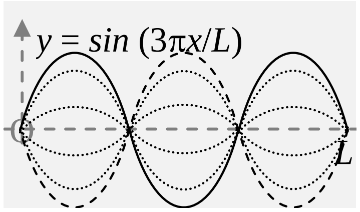


> Call the length of the string L. One possible solution gives the vertical position y along the string by
>
> $y \ = \ F_1(x, t) \ = \ \sin\,(\pi x/L)\cos\,(2\pi ft)$.
>
> The figure at right tries to suggest how it looks. The "standing wave" part [$\sin\,(\pi x/L)$] gives the enveloping arch. It is essential that the sine satisfies the **boundary condition**
>
> $F_1(0, t) \ = \ F_1(L, t) \ = \ 0 \qquad\qquad$ for all t.
>
> That corresponds to the string, as in a guitar, held fixed at both ends. The oscillating part [$\cos\,(2\pi ft)$] gives the vibration of the string at a frequency f. The frequency depends on the string's tension, length, and density (mass per unit of length).

(Match the statement about the oscillating part with experience. Increasing tension on a guitar string raises its pitch: Greater tension means greater force restoring the string to equilibrium (middle) position, therefore higher acceleration, quicker vibration, higher frequency. Increasing length does the opposite: On the left side of a piano or harp, longer wires mean weaker restoring force (because the tension, lying more nearly along the equilibrium line, has smaller perpendicular component), lower acceleration, slower vibration, lower frequency. Increasing density also lowers pitch: On all three instruments, the fat strings offer greater inertia, undergo smaller acceleration, slower vibration, lower frequency.)

> You might recognize, from the angle-sum formula, that the stated F_1 satisfies
>
> $2F_1(x, t) \quad = \quad \sin\,(x/[L/\pi] + t/[2\pi f\,]^{-1}) \quad + \quad \sin\,(x/[L/\pi] - t/[2\pi f\,]^{-1})$.
>
> That fits with d'Alembert's formulation (mentioned without detail in Section VIII.B.3a)
>
> $F(x, t) \quad = \quad f(x + t) \quad + \quad f(x - t)$
>
> for more general solutions.
>
> The possible solutions also include
>
> $y \ = \ F_2(x, t) \ = \ \sin\,(2\pi x/L)\cos\,(4\pi ft)$,
>
> $y \ = \ F_3(x, t) \ = \ \sin\,(3\pi x/L)\cos\,(6\pi ft)$,
>
> and so on. Those give the overlying shapes shown in the two figures at right. In those, the string is acting like two (or three or …) adjoining strings one-half (respectively one-third, …) as long as the original. Those produce tones at twice (three times, …) the original frequency.





The string equation has one of the most important and fruitful of mathematical properties, namely "linearity." Where the governing equation is linear, *superposition* applies. In less fancy language: If $F_1, F_2, \ldots, F_n$ are solutions, then so are the **linear combinations**

$a_1F_1 + a_2F_2 + \ldots + a_nF_n, \qquad\qquad$ with real constants $a_1, \ldots, a_n$.

Those are the combinations Bernoulli had in mind.

[In the sound of a string, F_1 is the string's "fundamental" vibration or "fundamental harmonic." The "overtones" $F_2, F_3, \ldots$ are "second harmonic," "third harmonic," …. Corresponding to their frequencies, the wavelengths are 1/2, 1/3, ... of the original wavelength. That's how the series 1 + 1/2 + 1/3 + … of reciprocals got its name.]

b) heat conduction

In 1807, Fourier presented a paper [titled] on the propagation of heat in solids. He introduced the PDE called the **heat conduction equation**. By analogy with Bernoulli, d'Alembert, and Euler on fluid flow, the equation describes heat flow by tracking the distribution of temperature in space (within a solid) and time. In a way, the equation extends Newton's law of cooling, incorporating conductivity (the speed of heat flow) and specific heat (resistance to temperature change). Interestingly, by then the idea of heat as a fluid flowing through objects had been discredited. The theory of heat as manifestation of molecular motion had begun to gain acceptance. Fourier's contemporaries were trying to describe heat transfer via interactions among discrete particles. Fourier instead treated the solid as a continuum.

His solutions extended Bernoulli's combinations to **trigonometric series**

$(a_1 \sin x + a_2 \sin 2x + \ldots) + (b_0 + b_1 \cos x + b_2 \cos 2x + \ldots)$.

There, the **Fourier coefficients** $a_1, a_2, \ldots$ and $b_0, b_1, \ldots$ are determined by the **boundary conditions**, the temperature distribution at $t = 0$ on the surface of the solid. The success of these series in solving the conduction equation was the ultimate trigger to reform of the way the math world treated infinite series. Lagrange, one of the judges of the paper, never moved from his objections to Fourier series, based on the question of convergence (coming up. Remember Lagrange's faith in Taylor series, for which Lagrange's own remainder (section VIII.B.4b) gives an estimate of the error of the series. No similar estimate was available with Fourier series.)

c) Fourier series

Even more problematic (than convergence) was Fourier's argument that *every* reasonable function is given by one of his trigonometric series.

> Take the most convenient example,
>
> $f(x) = x, \qquad 0 \le x \le 2\pi.$
>
> (You could work in any interval $a \le x \le b$, but then you would need to use sines and cosines of $2\pi x/(b - a)$, to make one complete sine or cosine wave fit exactly into the interval.)
>
> Fourier defined the sine coefficients by
>
> $a_n = (1/\pi) \int x \sin nx \; dx \qquad$ from $x = 0$ to $x = 2\pi$.
>
> The antiderivative of
>
> $g(x) = x \sin nx \qquad$ is $\qquad G(x) = (\sin nx)/n^2 - x(\cos nx)/n.$
>
> The change in G is
>
> $G(2\pi) - G(0) = -2\pi/n. \qquad$ (Last two statements are Exercise 1.)
>
> Therefore $a_n = -2/n$.
>
> For the cosine coefficients, the zero'th one has a different multiplier,
>
> $b_0 = (1/2\pi) \int x \cos 0x \; dx = (1/2\pi)(2\pi)^2/2 = \pi.$
>
> Otherwise,
>
> $b_n = (1/\pi) \int x \cos nx \; dx.$
>
> We omit the evaluations; see Exercise 2. All of these turn out to be zero.
>
> Fourier said that
>
> $$\begin{aligned} f(x) &= (a_1 \sin x + a_2 \sin 2x + \ldots) &&+ (b_0 + b_1 \cos x + b_2 \cos 2x + \ldots) \\ &= (-2/1 \sin x - 2/2 \sin 2x - 2/3 \sin 3x - \ldots) &&+ \pi. \end{aligned}$$

The cosine numbers are not a coincidence. If a function is odd relative to the halfway point, then its Fourier cosine coefficients after the one for [cos $0x$] are all zero. "Odd relative to the halfway point" means

$f(\pi + t) - f(\pi) = -[f(\pi - t) - f(\pi)].$

More simply, it means the graph of f is symmetric about the point $(\pi, f(\pi))$. If the function is even relative to the midpoint—if

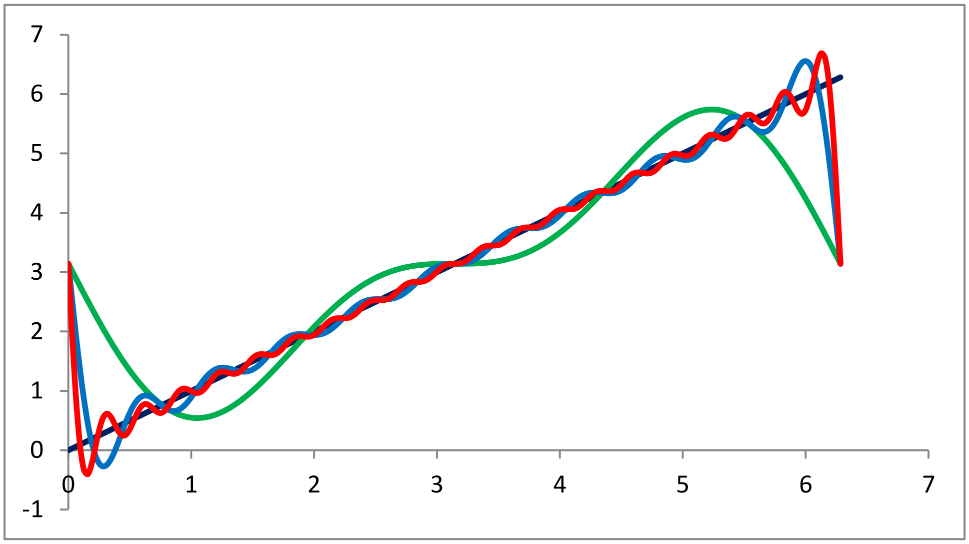
$f(\pi + t) = f(\pi - t),$

so that the graph is symmetric about the *line* $x = \pi$—then its Fourier sine coefficients are all zero. (See Exercise 3. Keep in mind that a function might be neither odd nor even.)

> You can show a corresponding behavior in Taylor series. Thus, sin x is **odd** (sin $(-x)$ = -sin x), and its Taylor series
>
> $\sin x \;=\; x - x^3/3! + x^5/5! - \ldots$
>
> has only odd-power terms. By contrast, cos x is even (cos $-x$ = cos x), and
>
> $\cos x \;=\; 1 - x^2/2! + x^4/4! - \ldots$
>
> has only even terms.

To understand the meaning and limitations of the series, look at the "partial sums"

$f_n(x) \;=\; \pi \;- 2/1 \sin x - 2/2 \sin 2x - \ldots - 2/n \sin nx.$



In the picture at right, we see the graph of f (black diagonal line), together with those of f_2 (green), f_{10} (blue), f_{20} (red). You can see that the sums coil increasingly tightly around the graph of f, except near the two ends.

> The difference at one or both ends is unavoidable. Each f_n has period 2π, whereas f is not periodic.
>
> From roughly $x = 0.7$ to $x = 5.5$, f_2 is a poor approximation to f, missing by up to 0.61. Over the same interval, f_{10} misses by only 0.21, f_{20} by only 0.13. On the bigger interval $x = 0.45$ to $x = 5.75$, f_{10} still lies within 0.31 of f. On the yet bigger $x = 0.35$ to $x = 5.90$, f_{20} is within 0.22 of f. The complete series matches f at every point of the interval other than the two ends.

Fourier managed to refine his arguments by the 1822 publication of *Theorie Analytique de la Chaleur* (… *of Heat*). From there, at least Cauchy was convinced that the theory was correct. Still, it took fifteen years before Peter Dirichlet gave the first acceptable proof that a function does equal its Fourier series. (That was under some mild restrictions. Those included the requirement that the function be periodic.)

Exercises IX.B.2

1. a) Use calculus to find the antiderivatives of $x \sin nx$.
   b) We have come to accept that if t is infinitesimal, then
   $[\sin (u + t) - \sin u]/t \;=\; \cos u$ and $[\cos (u + t) - \cos u]/t \;=\; -\sin u.$
   Use those relations to show, without using calculus, that
   $G(x) = (\sin nx)/n^2 - x (\cos nx)/n$
   has derivative
   $G'(x) = x \sin nx.$
   c) Evaluate the integral of ($x \sin nx$) between $x = 0$ and $x = 2\pi$.
2. a) Show with or without calculus that the antiderivative of
   $h(x) = x \cos nx$ is $H(x) = x (\sin nx)/n + (\cos nx)/n^2.$
   b) Evaluate the integral of ($x \cos nx$) between $x = 0$ and $x = 2\pi$.

3. Argue (via calculus or simply graphs) why, if a function F is odd relative to $x = \pi$, then its Fourier cosine coefficients

$$b_n = (1/\pi) \int F(x) \cos nx \; dx \qquad \text{from } x = 0 \text{ to } x = 2\pi, \qquad n \geq 1,$$

are zero; and similarly for the sine coefficients if F is even relative to $x = \pi$.

d) the importance of Fourier series

Return to our example with $f(x) = x$. It has

$$f(x) = (-2/1 \sin x + -2/2 \sin 2x + \ldots) + (\pi + 0 \cos x + 0 \cos 2x + \ldots) \qquad \text{for } 0 < x < 2\pi.$$

The series and the double sequence

$$(-2/1, -2/2, \ldots), \qquad (\pi, 0, 0, \ldots)$$

are both called the **Fourier transform** of *f*. Either encapsulates the continuum of data that *f* is, in a package of discrete elements. You can say that the series is the “spectrum” of *f*. Think of the function as the sound from an orchestra. Just as a prism separates sunlight into its constituent colors, so the transform separates *f* into its constituent wavelengths. Thus, $f(x)$ mixes a portion -2/1 of the signal $\sin x$ (wavelength π), -2/2 of $\sin 2x$ (wavelength $2\pi/2$), …; and it adds no portion of any signals $\cos nx$, except portion π of the steady tone with $n = 0$. [How would you characterize the last: infinite wavelength? Also, if you find it suspicious to add -2/1 of $\sin x$, just add 2/1 of $\sin (x + \pi)$.] Alternatively, and in contemporary language, you can say that the sequence *digitizes* the function.

Either of those interpretations helps explain why the series and the transform became important tools in solving the differential equations of mathematical physics.

On a practical level, return to the last figure (in (c)). While

$$(-2/1, -2/2, \ldots), \qquad (\pi, 0, 0, \ldots)$$

carries all the data of f, you can get most of the information from a finite subsequence. We saw that

$$f_2(x) = \pi -2/1 \sin x - 2/2 \sin 2x$$

is just a vague approximation to *f*. On the other hand, the graph of

$$f_{20}(x) = \pi -2/1 \sin x - 2/2 \sin 2x - \ldots - 2/20 \sin 20x$$

hugs most of the graph of *f*. For more than 80% of the interval, we can approximate the continuum of information in $f(x)$ with just the twenty-one numbers

$$(-2/1, -2/2, \ldots, -2/20), (\pi).$$

This idea of squeezing most of an infinity of information into a finite, and not too big, set of data underlies many “compression” schemes, essential to certain uses of computing. To transmit images from spacecraft, or create images from “magnetic resonance” signals, you typically have to take a partial Fourier transform to reduce some vast dataset to what machines can conveniently chew. The processes called MP3 and JPEG do that to make sound and images, respectively, compact and (you hope) faithful to the originals.

[Read T. N. Narasimhan’s wonderful article about Fourier’s life and influence, plus the relationship of his work to that of brilliant predecessors, like Lavoisier, and successors, like Kelvin and Einstein.]

3. Cauchy and Infinitesimal Analysis

We could have called this chapter “The Age of Rigor,” because that is what Cauchy ushered in. “Rigor” in mathematics refers to precise definitions and careful attention to axiomatic structure and logic. Its purpose is to avoid such traps as Euclid’s hidden assumptions (section III.A.5b) and the ambiguities of infinitesimals (what Berkeley criticized, section VIII.B.3a). We saw Cauchy bring it to the study of permutations, spanning roughly 1815-1844. Now we see how he brought rigor to calculus.

a) limits, continuity, derivatives

(i) limits

Cauchy resolved the infinitesimals paradoxes by elaborating on d'Alembert's notion of limits. (Remember that Newton and others had touched on the idea.) The resolution appeared in Cauchy's 1821 book *Cours d'Analyse*. (That became a well-used title, as "The Elements" had been in Greek times.)

Definition of Limit. The number L **is the limit of the function** f **at the place** $x = a$ if the values $f(x)$ can be forced to "[differ from L] by as little as one could wish" via keeping x *correspondingly* close to a.

There are multiple synonymous usages, including: "f approaches L" or "f tends to L" or $f \to L$; and "limit as x approaches a" or "as x tends to a" or as $x \to a$.

> Take an elementary example. It is natural for us to say that if $x \approx 2$, then $x^2 \approx 4$. In limit language, we would say that the limit of x^2 as x approaches 2 is 4. There is, of course, a notation:
>
> $$\lim_{x \to 2} x^2 = 4.$$
>
> To justify the statement, name a tiny distance, like 10^{-6}. Allowing x^2 to be on either side of 4, we take the absolute value $|x^2 - 4|$ of the difference. We want that to be smaller than 10^{-6}. By a property of absolute values,
>
> $$|x^2 - 4| = |(x + 2)(x - 2)| = |x + 2||x - 2|.$$
>
> Suppose we first confine x to the interval from 1 to 3. Then $x + 2$ is between 3 and 5, so that certainly $|x + 2| \le 5$. That means
>
> $$|x^2 - 4| \le 5|x - 2|.$$
>
> We can make the left side less than 10^{-6} by making the right side that small. Thus, the *corresponding* requirement on x is
>
> $$|x - 2| < 10^{-6}/5 = 0.000\,000\,2.$$
>
> Among the x's between 1 and 3, the ones we can guarantee to put x^2 within 10^{-6} of 4 are those between 1.999 999 8 and 2.000 000 2.

It is essential to see that in the last paragraph and the definition, there was no mention of the value of f *at the actual place* $x = a$. For the question of limit, that value need not exist. Above, x^2 has value 4 when x *is* 2. That value is irrelevant. What interests us is the range of values *in the vicinity* of $x = 2$.

> For $g(x) = (\sin x)/x$, $g(0)$ is not defined. Still, we have already given evidence that
>
> $$\lim_{x \to 0} g(x) = 1.$$
>
> Specifically, we argued in section VII.B.3b(i) that as long as $x \neq 0$,
>
> $$1/(2 - \cos x) < (\sin x)/x < 1.$$
>
> Therefore to keep $(\sin x)/x$ within say 10^{-6} of 1, it suffices to make
>
> $$0.999\,999 < 1/(2 - \cos x), \qquad \text{or} \qquad \cos x > 2 - 1/0.999\,999.$$
>
> Do Exercise 1 to decide how close to zero x has to be. See also Exercise 2.

Exercises IX.B.3a(i)

1. To determine the values of x near zero (on either side) for which

   $$\cos x > 2 - 1/0.999999:$$

   a) Use a scientific calculator to decide what x are needed.

   b) Skip calculation and use the acute-angle relation

   $$\cos x + x > \cos x + \sin x > 1$$

   to decide what x will *suffice*. (Observe how much territory the simplification costs you.)

2. a) Use the series

   $\sin x = x - x^3/3! + x^5/5! + \ldots$

   to confirm that $(\sin x)/x \to 1$ as $x \to 0$.

   b) It would be circular reasoning to say that (a) *proves* that $(\sin x)/x \to 1$ as $x \to 0$. Why?

(ii) no limits

Our most important functions—polynomials, sines and cosines, exponentials—always have limits (Exercise 1). It is important to see examples in which elementary functions fail to have limits.

A function can oscillate near $x = a$, without approaching a fixed value. Put

$F(x) = \cos(1/x)$, for $x \neq 0$.

Consider the place $a = 0$, where we have deliberately left F without a value. At the nearby places

$x = 1/([10^6 + 1]\pi)$, $1/([10^6 + 2]\pi)$, $1/([10^6 + 3]\pi)$, …,

the values of F are -1, 1, -1, 1, …. You cannot force them all to lie within 0.1 of any possible L.

More familiar is the situation where the values increase beyond bound. Call

$G(x) = 1/x^2$, for $x \neq 0$.

Here, the places

$x = 10^{-6}$, 10^{-9}, 10^{-12}, …

are all close to zero, but the values $G(x)$ are far apart; they cannot be close to any particular L. In this case, we can salvage some information: We may choose to say that the limit of G "is infinity" (Exercise 2).

For one last example, a function can have different tendencies on the two sides of $x = a$. Set

$H(x) = |\sin x|/x$, for $x \neq 0$.

This H matches $(\sin x)/x$ for $x > 0$. Therefore it approaches 1 as x approaches zero from the positive side. For $x < 0$, though, H matches $-(\sin x)/x$. That means $H(x) \to -1$ as $x \to 0$ from the negative side. Again, we may choose to say that the limit **from the right** (or **left**) is 1 (respectively -1). (Sometimes we have no choice: We can only discuss $h(x) = \sqrt{x}$ from the right.)

These limit-less examples reflect an observation Cauchy made. For f to have a limit L someplace, the nearby values of f have to be close to L. In that case, the nearby values *are close to one another*. The observation led to what we call "Cauchy's criterion."

Proposition. (Cauchy's Criterion) The function f has a limit at $x = a$ iff pairs of values $f(s)$, $f(t)$ can be forced to differ by as little as desired by keeping s and t correspondingly close to a.

It is fairly clear why (limit exists) implies (close values). The converse is harder, and we skip it. The important thing to note about the criterion is that it does not mention, and cannot figure out, the limit (if there is one). However, as we saw with $1/x^2$ and $\cos(1/x)$ at $x = 0$, it can guarantee that there is no limit.

The most extreme limit-less example came from (Peter Gustav Lejeune) Dirichlet: He produced a function that *never* has a limit. Dirichlet made many contributions to mathematics; read about them from J J O'Connor and E F Robertson at St Andrews. [Two are as elementary as they are well-known. One settles Fermat's last theorem for exponents 5 and 14. The other is often called "Dirichlet's theorem." It says that in any arithmetic progression in which the first term is relatively prime to the constant increment—in 4, 13, 22, 31, … just as in 1, 2, 3, ... —there exists an infinity of primes.] We already mentioned that Dirichlet, who met Fourier (and eventually succeeded Gauss at Göttingen), gave conditions to guarantee that a function's Fourier series adds up to the function. To do that, he needed Cauchy's definition of series adding up to something (still coming up). More fundamentally, he needed to establish what a function is.

> Dirichlet’s definition is practically what we use today: A **function** is given by a rule (could be a formula, could be some complicated set of instructions) that assigns to each member of a chosen set (the **domain**) a unique member of a second set (the **range**). Our old cubic substitution
>
> $$x = u + 27/u$$
>
> serves to give x as a function of u, because for a given nonzero u, it forces exactly one value of x. It does not give u as a function of x, because for $x = 12$, it allows the two values $u = 9$ and $u = 3$.
>
> The rule for what we call “Dirichlet’s function” is
>
> $$\begin{array}{lcll} D(x) & = & 1 & \text{if } x \text{ is a rational real number,} \\ & = & 0 & \text{if instead } x \text{ is irrational.} \end{array}$$
>
> This function is the opposite of our favorites. It fails Cauchy’s criterion everywhere. At x =1, for example, the rational $s = 1 + 10^{-100}$ and irrational $t = \sqrt{}(1 + 10^{-100})$ are exceedingly close (to 1 and to each other), but $D(s)$ and $D(t)$ do not “differ by as little as desired.” (Do Exercise 3.)

(iii) continuous functions

With limits in hand, Cauchy defined continuity. We say that f **is continuous at** $x = a$ if

$$\lim_{x\to a} f(x) \ = \ f(a).$$

That last clause is simple (not compound), but it makes a triple demand. It requires that f have a value $f(a)$, that it have a limit as $x \to a$, and that the value match the limit. On the evidence of Exercise 1, we see that our important functions are continuous everywhere.

Calculus had gotten along well enough without defining function, let alone what it means for one to be continuous. The intuitive notion of function simply used formulas. Continuity was decided by their graphs. If the graph had no breaks—if you could sketch it without picking the pencil off the paper—then the function was continuous. That notion of continuity was helpful, but it leads to paradoxes. You can have a function whose graph has no breaks, yet you cannot (even theoretically) draw all of it in one move (Exercise 4a-c).

Forget the possible deficiencies in the intuitive notion. Observe instead that Cauchy’s definition really does advance the rigor. It turns continuity into a quantitative notion. Lagrange would say it renders the concept *analytical*, meaning algebraic (section VIII.B.4a); Wallis would say it *arithmetizes* continuity (section VII.A.6).

(iv) derivatives

Finally, Cauchy arrived where infinitesimals had operated. Recall Fermat’s way:

> Take say $f(x) = x^3$. Look at the slope of (what we now call) the secant joining the points
>
> $$(a, f(a)) \qquad \text{and} \qquad (a + h, f(a + h)),$$
>
> namely
>
> $$[f(a + h) – f(a)]/h \ = \ 3a^2 + 3ah + h^2.$$
>
> Then set $h = 0$.

Cauchy overcame the objection to that last step by defining

$$f'(a) \ = \ \lim_{h\to 0} \ \big([f(a \ + \ h) - f(a)]/h\big).$$

That limit is $3a^2$, based on Exercise 1a, because we treat $3a^2 + 3ah + h^2$ *as a polynomial in h*. (Cauchy used “variable quantity” in place of “function.” When many varying quantities (x, a, f, and h) are involved, the name is clearly a good idea.)

Whenever the limit exists, we say f **is differentiable at** or f **has a derivative** at $x = a$, and call the limit the **derivative** of f there. For our important functions, finding the limit of the slope of the secant is direct and gives the derivatives we already know. Our favorites are differentiable everywhere.

We are deliberately skipping the precise definition of integral. It is the limit of a sum of terms—actual numbers, not infinitesimals—but it is not worth our time to set it up here. Again, it is not hard to show that the limit exists for our functions and to find the limit. In fact, every continuous function has an integral. However, proving that statement requires discoveries that are at least thirty years ahead of us.

The statement that our important functions are continuous and differentiable brings up an important point. For a function to be differentiable, it first has to be continuous.

> Suppose
> $$[f(a+h)-f(a)]/h$$
> has limit $L = f'(a)$. Then
> $$f(a+h)-f(a) \approx hL \approx 0 \qquad \text{as } h \to 0.$$
> That says $f(a+h) \to f(a)$. It implies that $f(a)$ is the limit of $f(x)$, making f continuous.

The converse is false. A function can be continuous at $x = a$ and have $f'(a)$ undefined (Exercise 4d).

Exercises IX.B.3a(iv)

1. Take our example
   $$\lim_{x\to 2} x^2 = 2^2$$
   as evidence that for any integer $n \geq 0$ and any a,
   $$\lim_{x\to a} x^n = a^n.$$
   Give evidence that
   a) If p is a polynomial, then
   $$\lim_{x\to 10} p(x) = p(10).$$
   b) $\lim_{x\to a} \sin x = \sin a$. (Hint:
   $\sin x - \sin a = \sin([x + a]/2 + [x - a]/2) - \sin([x + a]/2 - [x - a]/2)$.
   c) $\lim_{x\to a} \cos x = \cos a$. (Hint: Adapt the previous hint.)
   d) $\lim_{x\to a} e^x = e^a$.
2. We can bring rigor to Wallis's statement that "1/0 is infinity" with this definition:
   The limit of *G* as *x* approaches *a* **is infinity** if the values $G(x)$ can be forced to *exceed any number we wish* by keeping *x* correspondingly close to *a*.
   Show how the values of $G(x) = 1/x^2$ around $x = 0$ can be forced to exceed 10^{100}.
3. a) How far from the numbers 1 and $s = 1 + 10^{-100}$ is the number $t = \sqrt{s}$?
   b) Show is *t* irrational.
   c) For Dirichlet's function *D*, how much is $D(s) - D(t)$?
4. Set
   $$G(x) = x\cos(1/x) \quad \text{when } x \neq 0, \qquad G(0) = 0.$$
   a) Make a sketch suggesting the graph of $G(x)$, $0 \leq x \leq 3/\pi$. (Hint: Spot the places where $G(x)$ matches either x or $-x$, and the intermediate places where $G = 0$.)
   b) Based on your picture, why is *G* continuous at every *x* from 0 to $3/\pi$? (Notice that it was essential to specify a value $G(0)$.)
   c) From the picture, why is it impossible to draw the complete graph from end to end?
   d) Show that *G* is not differentiable at $x = 0$. (Hint: No matter how tightly you restrict *h* near 0, the slopes of the secants from $(0, G(0))$ to $(0 + h, G(0 + h))$ can range from 1 to -1; there is no limit to the slopes of the secants.)

5. The **greatest integer function** or **floor function** is defined by the special bracket
   $\lfloor x \rfloor$ = the biggest integer that does not exceed x.
   (Thus, $\lfloor 3/2 \rfloor = 1$, $\lfloor -3/2 \rfloor = -2$, and $\lfloor -2 \rfloor = -2$.) At what places is $\lfloor x \rfloor$ continuous?

b) series

Cauchy turned to limits again to give precise meaning to series.

From the series

$$a_1 + a_2 + a_3 + \ldots,$$

write the **partial sum**

$$s_n = a_1 + a_2 + \ldots + a_n.$$

If the partial sums approach some real number s, then we attach the value s to the series. We need to specify in what sense they "approach." Wallis would have said, when $n = \infty$. Cauchy said, as n approaches infinity. That brings the question down to defining "n approaches infinity."

Definition of Series (Convergence). The series $a_1 + a_2 + a_3 + \ldots$ **converges** to the real number s if we can force the partial sums s_n to differ from s by as little as we wish by keeping *n correspondingly* large.

If the series converges to s, we call s its **sum** or **value**. For convergence, the definition demands that eventually the partial sums get close to s and *stay close*.

Immediately, we say goodbye to

$$1 - 1 + 1 - 1 + \ldots.$$

The partial sums are 1, 0, 1, 0, …. There is no real s for which the partial sums get and stay within even 0.4. The series **diverges** (does not converge), and we may not speak of its value.

Even the Egyptians could sum a geometric series, like

$$1 + 1/3 + 1/3^2 + 1/3^3 + \ldots.$$

We know from multiple sources that

$$\begin{aligned} s_n = 1 + 1/3 + \ldots + 1/3^{n-1} &= (1 - 1/3^n)/(1 - 1/3) \\ &= 3/2 - (3/2)/3^n. \end{aligned}$$

To keep s_n within say 10^{-10} of 3/2, we just need

$$(3/2)/3^n < 10^{-10}, \qquad \text{or} \qquad 3^n > 1.5 \times 10^{10}.$$

Since $3^3 > 1.5 \times 10$, we have $3^{30} > 1.5^{10} \times 10^{10}$. For at least $n \geq 30$, s_n is between $3/2 - 10^{-10}$ and 3/2.

Under the definition, the harmonic series does not converge. (Reason?) Still, it makes sense to say it has a "value," namely infinity. Accordingly, we will allow ourselves to say it "converges to infinity" (instead of "diverges to infinity"), under a definition that is Exercise 2.

Naturally, Cauchy's observation applies to these limits, too. If the partial sums are close to s, then they are close to each other.

Proposition (Cauchy's Criterion). A series converges to a real sum iff we can force pairs s_m and s_n of its partial sums to differ by as little as we wish, by keeping both m and n correspondingly large.

The rule is not complicated. Observe that if $m > n$, then the distance from s_m to s_n is

$$\begin{aligned} |\, s_m - s_n \,| &= |\, (a_1 + a_2 + \ldots + a_m) - (a_1 + a_2 + \ldots + a_n) \,| \\ &= |\, a_{n+1} + a_{n+2} + \ldots + a_m \,|. \end{aligned}$$

It is the absolute value of the sum of a finite bunch of consecutive terms. The criterion demands that such sums get small and stay small.

Recall Jacques Bernoulli's consideration of

$$1/1^2 + 1/2^2 + 1/3^2 + \ldots.$$

It is hard to relate to the sum Euler's screwball argument found, but easy to relate to the criterion.

Modifying Jacques's observation, we have

$$1/1001^2 + 1/1002^2 + \ldots + 1/(10^{60})^2 < 1/1000(1001) + 1/1001(1002) + \ldots + 1/(10^{60} - 1)10^{60}.$$

Here, we are not comparing series. Those are actual sums, albeit of a lot of terms. In each term on the left, the denominator exceeds its cousin on the right. Therefore the sum on the left is less than the one on the right. The latter is easy to evaluate: It is

$$[1/1000 - 1/1001] + [1/1001 - 1/1002] + \ldots + [1/(10^{60} - 1) - 1/10^{60}]$$
$$= \quad 1/1000 - 1/10^{60}. \qquad \text{(Why, again?)}$$

The sum of any string of consecutive terms starting with $1/(n + 1)^2$, no matter how many terms you sum, is less than $1/n$. By Cauchy's criterion, the series converges. (Note, incidentally, that the same is true for $1/(1 \times 2) + 1/(2 \times 3) + 1/(3 \times 4) + \ldots$, in agreement with Exercise 1.)

Another interesting example is the **alternating harmonic series**

$$1 - 1/2 + 1/3 - 1/4 + \ldots.$$

If you start the string at a positive term, like 1/1001, you find

$$0 \quad < \quad 1/1001 - 1/1002 + \ldots \pm 1/m \quad < \quad 1/1001.$$

Just pair up the terms this way,

$$(1/1001 - 1/1002) + (1/1003 - 1/1004) + \ldots + (1/[10^{60} - 1] - 1/10^{60}),$$

to see that the sum is positive, with or without the red term. Pair them this way,

$$1/1001 - (1/1002 - 1/1003) - (1/1004 - 1/1005) - \ldots - (1/[10^{60} - 2] - 1/[10^{60} - 1]) - 1/10^{60},$$

and you see that the sum is less than 1/1001. If instead you start a string at a negative term, you find

$$-1/998 < \quad -1/998 + 1/999 + \ldots + 1/[10^{60} - 1] - 1/10^{60} \quad < \quad 0 \qquad \text{(Exercise 4).}$$

Either way,

$$0 \quad < \quad |\pm(1/[n + 1] - 1/[n + 2] + \ldots \pm 1/m)| \quad < \quad 1/[n + 1].$$

The sums of consecutive terms get small.

Check in the previous paragraph's argument that three facts are the keys to convergence:

1. The series is alternating. The sequence of signs is +, –, +, –, ….
2. The absolute values decrease. Forgetting the signs, $1/[n + 1] < 1/n$.
3. The terms approach zero as n approaches infinity. (Remember how those "approaches" are defined.)

If a series has those three properties, then the argument shows that the series satisfies Cauchy's criterion, is therefore convergent. That principle (criterion?) is called the **alternating series test** (or rarely, "Leibniz's test," after its discoverer.)

Finally, the condition #3 above is necessary for *any* series to converge to a real. We said that for convergence, the sum of any string of consecutive terms has to get small. Such strings include single terms. For $a_1 + a_2 + a_3 + \ldots$ to converge, it is necessary that $a_n \to 0$ as $n \to \infty$. At the same time, the converse if false: The series might diverge *even if* $a_n \to 0$. (Give an example.)

Exercises IX.B.3b

1. Find the value of

   $$1/(1 \times 2) + 1/(2 \times 3) + 1/(3 \times 4) + \ldots,$$

   and show that your value satisfies the definition.

2. a) Define what it means for $a_1 + a_2 + a_3 + \ldots$ to **converge to infinity**. (Hint: Use Exercise IX.B.3a(iv):2 as a model.)

   b) Prove that the harmonic series satisfies the definition in (a).

   c) Does the series

   $$1 - 2 + 3 - 4 + \ldots$$

   converge? Does it converge to infinity?

3. Use Cauchy's criterion to prove that each series converges:
   a) $10^6 + (10^6)^2/2! + (10^6)^3/3! + \ldots.$ (What is its value?)
   b) $10^6 – (10^6)^3/3! + (10^6)^5/5! – \ldots.$ (Hint: If (a) meets the criterion, then (b) has to.)
4. Show that with or without the red term,
   $$-1/998 \quad < \quad -1/998 + 1/999 + \ldots + 1/[10^{60} – 1] – 1/10^{60} \quad < \quad 0.$$
5. Show that the partial sum
   $1 – 1/2 + 1/3 – 1/4 + \ldots \pm 1/n$ (whichever sign is right)
   misses the value of the alternating harmonic series by no more than $1/(n + 1)$. (That works whenever the alternating series test applies: The error is no more than the next term.)
6. Argue why:
   a) $1/\sqrt{1} + 1/\sqrt{2} + 1/\sqrt{3} + 1/\sqrt{4} + \ldots$ converges to infinity.
   b) $1/\sqrt{1} – 1/\sqrt{2} + 1/\sqrt{3} – 1/\sqrt{4} + \ldots$ converges.
   c) $1/(1 + 2) – 2/(2 + 3) + 3/(3 + 4) – 4/(4 + 5) + \ldots$
   does not converge to either a real number or to infinity.
7. Use the example
   $$(5 + 2)^{4/3} = 5^{4/3} + (4/3)\, 5^{1/3}2 + (4/3)(1/3)/2!\, 5^{-2/3}2^2 + (4/3)(1/3)(-2/3)/3!\, 5^{-5/3}2^3 + \ldots$$
   to give evidence that the binomial series converges when the binomial's second term has smaller absolute value than the first.

4. Bolzano, Dedekind, and the Nature of the Real Numbers

The axiomatization of the calculus evolved over more than fifty years. One of the key elements was precise definition of the real numbers. Here we will see some of the evolution.

a) the ordered-field properties

There are three properties that together characterize the reals. The first is that the set **R**, with its addition and multiplication, constitutes a field. The definition of field is in section IX.A.2c(ii). There you will also find statements—some proved, some left as exercises, some merely talk—to the effect that the algebraic manipulations we work on expressions or equations are consequences of the field axioms. For some examples, subtraction and division are definable, $a – b + c$ and $(a/b)/(c/d)$ make sense, and

$$a – b + c = a – (b – c) \qquad \text{and} \qquad (a/b)/(c/d) = (ad)/(bc).$$

The second property is that the field of reals has an order. Here we are departing from the use of "order" in the sense of say

$$3^6 \equiv 1 \qquad \text{modulo } 7.$$

We will use the word in the sense of "*a* comes after *b*," so that we speak of *a* being larger than *b*.

The Order Axiom. In a field *K*, **an order** is a relation ">" that obeys three rules:
Rule 1. It is compatible with the addition. Specifically, for elements *a*, *b*, *c*, *d* in *K*:

If $a > b$ and $c > d$, then $a + c > b + c$ and $a + c > b + d$.

Rule 2. It is compatible with the multiplication. Specifically, for elements *a*, *b*, *c* in *K*:

If $a > b$ and $c > 0$, then $ac > bc$.

Rule 3. (The **Trichotomy**) It relates different elements one way or the other. That is, if *a* and *b* are elements of *K*, then *exactly one* of the following is true:

$a > b$, or $a = b$, or $b > a$.

We read "$a > b$" as you expect, "*a* is bigger than *b*." We may write "$b < a$" ("*b* is less than *a*") to mean the same thing. We also write "$a \geq b$" to encompass the two mutually exclusive possibilities $a > b$ and $a = b$, and similarly with $b \leq a$. Immediately, we have a result.

Theorem 1. $1 > 0$.

[This is one of my favorite statements, because it appears to say something we learn by age 3. By then, we know that one cookie is better than zero cookies, as surely as two cookies are better than one cookie. But the statement has nothing to do with cookies, or even with quantities. It is a technical mathematical statement. It says that in an ordered field, the multiplication identity bears the order relation to the addition identity. As such, it has to be addressed via the axioms.]

> The proof is simple. Rule 3 allows just three possibilities; we simply rule out the later two. The middle one is out: $1 = 0$ is not allowed in a field (section IX.A.2c(ii)). The third one leads to a contradiction. Suppose $0 > 1$. By Rule 1,
>
> $$0 + -1 \ > \ 1 + -1.$$
>
> That says $-1 > 0$. By Rule 2,
>
> $$(-1)(-1) \ > \ 0(-1).$$
>
> That says $1 > 0$. The contradiction lies there; the trichotomy does not allow $1 > 0$ at the same time as $0 > 1$. With $1 = 0$ and $0 > 1$ disallowed, the order has to be $1 > 0$.

When $a > 0$, we say a **is positive**. If $0 > a$, then a is **negative**. [That's the reason for preferring to read "$-a$" as "minus a," rather than "negative a." An additive inverse might be positive, like $-(-1)$.] That leaves 0 by itself, neither positive nor negative.

We can now say that most of the manipulations we do with inequalities follow from the field plus order axioms. We will prove an important one and leave others to exercises; view Exercise 1.

Theorem 2. In an ordered field, the order is transitive:

If $a > b$ and $b > c$, then $a > c$.

> Assume $a > b$ and $b > c$. By Exercise 1a, $a - b \ > \ 0$ and $b - c \ > \ 0$. By Rule 1,
>
> $$(a - b) + (b - c) \ > \ 0 + 0.$$
>
> The left side is $a - c$, the right side 0. Therefore $a - c \ > \ 0$, and $a > c$.

In view of the transitivity, we abbreviate ($a > b$ and $b > c$) by

$$a > b > c.$$

Based on the order axiom, we can distinguish the real field from some of the fields we named in section IX.A.2c(ii) . On one hand, **R** is bigger than all the finite fields. For a more fundamental difference, recall that the set $\mathbf{Z}_p$ of residues modulo a prime p constitutes a field under addition and multiplication modulo p. The next theorem implies that there is no way to put an order on it.

Theorem 3. Every ordered field is infinite. In fact, within any ordered field, there is a subfield that is a copy of the rational numbers.

> We know that a field has to have at least two elements, the two identities. In an ordered field, Theorem 1 says, $1 > 0$. Therefore by Rule 1,
>
> $$1 + 1 \ > \ 0 + 1 \ = \ 1.$$
>
> By transitivity, $1 + 1 \ > 0$. Consequently in any ordered field, 0, 1, and $(1 + 1)$ are three distinct members. Similarly,
>
> $0, 1, 1 + 1, 1 + 1 + 1, \ldots$ (henceforth 0, 1, "2," "3," …)
>
> is an endless list of unequal elements. Therefore the field is infinite, containing a copy of the natural numbers. The copy of the rationals is yours to find in Exercise 2.

On the other hand, **R** cannot be as big as the complex field **C**. The latter includes negative nonzero squares, like i^2 ; by Exercise 1c, nonzero squares in **R** have to be positive. Indeed, nonzero squares have to be positive in *any* ordered field. Accordingly in **C**, as in the finite fields, there cannot be an order.

By Theorem 3, the rationals constitute a subfield of **R**. (Is it all of **R**?) By the previous paragraph, **R** is a subfield of **C**, and not all of **C**. We have the reals bracketed between the rationals and the complexes. That leaves us with the question of what distinguishes **R** from **Q**. We answer it next.

Exercises IX.B.4a

1. Prove that for reals a, b, c:
   a) $a > b$ iff $a - b > 0$.
   b) If $a > b$ and $c < 0$, then $ac < bc$.
   c) If $a \neq 0$, then $a^2 > 0$.
   d) If a and b are nonzero, then $a/b > 0$ iff either ($a > 0$ and $b > 0$) or ($a < 0$ and $b < 0$).
2. Show that an ordered field contains a subfield identical to the rationals. (Hint: Follow the "subfields of **R**" argument in section IX.A.2c(iii).)
3. In an ordered field, the set of positive elements is not just a byproduct of the order. Its existence is actually equivalent to the existence of the order. Prove the equivalence:
   a) Assume a field is ordered. Show that its subset P of positive elements satisfies:
      (i) It is closed under addition; if a and b are positive, then so is $a + b$.
      (ii) It is closed under multiplication; if a and b are positive, then so is ab.
      (iii) (Trichotomy) It partitions the field into three disjoint sets; for every element a, exactly one of these is true: a is in P, or $a = 0$, or $-a$ is in P.
   b) Assume a field has a subset S that satisfies (i)-(iii). Show that the relation "$\rangle$" defined by
      $c \rangle d$ means $c - d$ is in S
   obeys Rules 1-3.

b) the continuum properties

The third property characterizes the ordered field of real numbers. It was discovered by a Bohemian (Czech), Bernard Bolzano (1781-1848). His work was completely unknown, perhaps because he was a priest and not a math professional, until Weierstrass (just ahead) rescued it from obscurity decades later. It is remarkable how it completes the picture of the reals. ("Completes" happens to be a technical, and most appropriate, term; see it at St Andrews.)

(i) Bolzano's axiom

The ancient notion of real number was geometric. It coincided with length. It was in terms of length that the Pythagoreans cast their proof that the rationals do not account for all "numbers." Around 1817, Bolzano was first to *arithmetize* the reals, to describe them in terms of numbers alone.

Bolzano's Axiom. Suppose there is a property that some, but not all, real numbers have. Assume that every real number with the property exceeds every real that lacks it. Then there is either a smallest real number with the property, or a largest real without it.

["Bolzano's axiom" is not a standard name, but it fits. Remember that it is an assumption, not something to be proved.]

"Some, but not all" is clumsy enough to suggest putting the axiom into the language of sets.

Bolzano's Axiom in terms of sets. Suppose you split the real numbers between nonempty sets L and R: $\mathbf{R} = L \cup R$, L and R not empty. Assume that each element of R is rightward of every element of L: (r is in R and l is in L) forces $r > l$. Then either R has a least element, or L has a biggest one.

(In either version: Can it be both—biggest l and smallest r? Do Exercise 1.)

Notice that we did not specify that L (as in "left") and R (right) need be disjoint, but every l's being leftward of each r forces it. Notice also "right" and "left"; we keep some of the language of geometry.

The best way to illustrate the meaning of "continuum" is a principle we have invoked before, strictly on the basis of intuition. It relates directly to solutions of equations.

Theorem 1. (The Intermediate-Value theorem) Suppose f is continuous everywhere from $x = a$ to $x = b$. Assume that $f(a)$ is different from $f(b)$. If d is a real number between $f(a)$ and $f(b)$ (an "intermediate value"), then there is a place c between a and b at which $f(c) = d$.

(Here we need to note that we wrote Cauchy's definitions of limit and continuity, plus Dirichlet's definition of function, ahead of this subsection's characterization of the reals. Review those definitions (section IX.B.3a) to see that they depend on just the ordered field properties. Thus, for example, to say that $f(x)$ is within 10^{-6} of 1 is to say that

$(10^6)^{-1} > f(x) - 1$ and simultaneously $(10^6)^{-1} > 1 - f(x)$.

Those are equivalent to

$1 + (10^6)^{-1} > f(x) > 1 - (10^6)^{-1}$.

It could be that f is undefined left of $x = a$ or right of $x = b$. In either case, "f is continuous" means that $f(a)$ matches the limit of f *from the right*, and similarly at b from the left.)

We illustrate the proof with an example. To use an old friend, recall Fibonacci's cubic equation

$x^3 + 2x^2 + 10x = 20$.

Call the left side $f(x)$, with $d = 20$. We have $f(1) = 13$ and $f(2) = 36$. We will show that somewhere between $x = 1$ and $x = 2$, there is a c where $f(c) = d$. (Try to see in the argument below that what we do is trace the graph of f, leftward from $x = 2$, until we hit a place where $f(x) \leq 20$.)

We know f is continuous for all x, and $f(1) < 20 < f(2)$. Look at the set R of real r with the property

$r \geq 2$, or instead $r < 2$ and from $x = r$ to $x = 2$, $f(x) > 20$.

Clearly 2 is in R and 1 is in the complement L. Suppose r is in R and l in L. Necessarily $l < 2$. It cannot be that $r \leq l < 2$; in that case, because f exceeds 20 at r and rightward, l would be in R. Therefore $r > l$, and the axiom applies.

There cannot be a smallest r in R. If r is in R, then $f(r) > 20$. Say $f(r) = 21$. By the definition of continuity, we can force $f(x)$ to stay within 1/2 of 21, meaning between 20 + 1/2 and 21 + 1/2, by keeping x between say $r - 1/1000$ and $r + 1/1000$. Hence at $r - 1/2000$ and rightward, $f(x) > 20$. Therefore $r - 1/2000$ is also in R, and r is not the smallest member.

Consequently there is a biggest l in L. All the places above l are in R. That means $l > 1$, because the continuity implies that $f(x)$ stays below 13 + 1/2 just to the right of 1. Because all the places above l are in R, $f(x)$ has to exceed 20 rightward of l, at least until $x = 2$. That means $f(l) > 20$ is not possible; that would mean l is in R. But $f(l) < 20$ is also impossible. If it were true, then again by the continuity argument, f would stay less than 20 to the immediate right of l. There remains only one possibility: $f(l) = 20$. At the place $c = l$ strictly between 1 and 2, f takes on the intermediate value.

From Bolzano's axiom, we find that the cubic has a real root. From Leonardo, we knew that the cubic has no rational root. If we did not realize it before, we know it now: The ordered field of real numbers includes members that are not rational, and the axiom points to where the difference originates.

Finally, it is useful to observe that modern physics departed from classical physics by ceasing to think of the universe in continuum terms. For example, early in the twentieth century, Niels Bohr explained the behavior of electrons by hypothesizing that their speed—more accurately, their energy—is "quantized." Classical physics would have thought a particle can have one unit of energy, or two units, or any intermediate value. Bohr said it can have a "ground state" energy or various levels of "excited" energy, *but not anything in between*. You can excite the electron by delivering to it a "quantum" of light, a “photon,” of exactly the energy needed to reach the next level, but not with less; and from that higher level the electron can give back that quantum of energy (and return to the ground level), but not less.

Exercises IX.B.4b(i)

1. Under the hypotheses of (the first version of) Bolzano’s axiom, show that there cannot exist both a smallest real with the property and a greatest real without it.
2. a) Prove that **R** includes a number whose square is 2.
   b) More generally, prove that if n is natural and $a > 0$ real, then a has a real n’th root.

(ii) the least upper bound property

There is a property equivalent to Bolzano’s axiom that is now the usual one used to characterize the reals. It has the advantage of not needing to partition the entire field.

The Least Upper Bound Property. If a nonempty subset of **R** has an upper bound, then it has a least upper bound.

We say the real number u is **an upper bound** for the set S (empty or not) if every s in S has $s \leq u$. If S has an upper bound (in which case it has a lot of them), we say S **is bounded above**. Let

$S = \{1, 1/2, 1/3, 1/4, \ldots\}$ and $T = \{1/2, 3/4, 7/8, \ldots\}$.

Clearly 1 is the largest member of S. In that situation, 1 is an upper bound, and nothing smaller is. (Why?) Hence 1 is the least upper bound (or **LUB** or **supremum**) of S. Equally clearly, no member of T is biggest. But all members of T are *below* 1, and some members exceed say $1 - 1/2^{50}$. Therefore among the upper bounds T has, 1 is least.

Theorem 2. In an ordered field, Bolzano’s axiom is equivalent to the LUB property.

Assume first Bolzano’s axiom. Suppose S is nonempty and is bounded above. Let R be the set of field elements that are upper bounds of S, L the rest. By assumption, R has some members. There is at least one s in S, which means that $s - 1$ is not an upper bound. Hence at least $s - 1$ belongs to L. If r is in R and l in L, then r is an upper bound of S and l is not. That means there exists some t in S with $l < t$, and necessarily $t \leq r$. Therefore $l < r$; Bolzano’s axiom applies.

There is no biggest l. That is, if l is not an upper bound of S, then there is an s in S with $l < s$. Necessarily, the average $(l + s)/2$ is also less than s, is therefore not an upper bound either, and exceeds l. Because there is no biggest l, there must be a smallest member of R. Among the upper bounds, there is a least one. That proves the LUB property.

Now assume the LUB property. Suppose R and L partition the field as the hypotheses of Bolzano’s axiom provide. Then L is nonempty and has as upper bounds all the members of R, of which there are some. By the property, L has an LUB; call it c. If c is in L, then it is the largest member of L, all the others being smaller. If c is *not* in L, then it is in R. In that case, any member r of R, being an upper bound for L, must have $r \geq c$. That says c is the smallest member of R. We infer that either L has a biggest member, or R has a smallest. That proves Bolzano’s axiom.

Each of Bolzano’s axiom and the LUB property (the two “continuum properties”) implies the other. That is what Theorem 2 says.

In section III.B.3, we noted that the “axiom of Archimedes,” which the Syracusan ascribed to Eudoxus, helps to complete our picture of the reals. That statement is a consequence of the description we have given here for **R**.

Theorem 3. If a and r are real numbers, with $a > 0$, then some multiple of a exceeds r.

The “multiples of a” are a, $2a$, $3a$, …. Here 2, 3, … are not natural numbers. We attached those names to the elements $1 + 1$, $1 + 1 + 1$, … of the ordered field of reals.

> The number r cannot exceed all of the multiples. If it did, then
> $S = \{a, 2a, 3a, \ldots\}$
> would have an upper bound. It would therefore have a least upper bound l. If l were the LUB, then $l - a$ would not be an upper bound. There would be a multiple $na > l - a$. That would make
> $(n + 1)a \;=\; na + a \;>\; l,$
> contrary to l being an upper bound. It follows that one of the multiples must exceed r.

Look to Exercise 1 for more of "our picture of the reals."

Exercises IX.B.4b(ii)

1. Show that in an ordered field with the LUB property:
   a) If r is positive, then there exists $n = 1 + 1 + \ldots + 1$ such that
   $0 \;< 1/n \;<\; r.$
   b) Between any two members, there exists a rational m/n.
   c) Between any two members, there exists an infinity of rationals.
   d) Between any two members, there exists an infinity of irrationals.
2. Show that in the same field, every polynomial of odd degree has a root.
3. Show that this statement is equivalent to Bolzano's axiom:
   Suppose that R is a nonempty subset of **R**. Assume that there is a real u such that every real $l < u$ is outside R. Then there is a biggest such u: a place U to whose left every real is outside R, but to whose right no place V has everything to *its* left outside R. (This is a set-based version of what **Kline**, page 953, writes in terms of properties. Kline notes that this version establishes the least upper bound property.)
4. We say the real number v is **a lower bound** for the set S if every s in S has $s \geq v$. Show that the LUB property is equivalent to the **GLB property**: If a nonempty set T has a lower bound, then it has a greatest lower bound.

c) Dedekind's realization

(Julius Wilhelm) Richard Dedekind (1831-1916) was a student of Gauss, and later worked with Dirichlet. He made multiple advances toward setting precise definitions for our familiar number systems. The progress went from the natural numbers to the integers, to the rationals, to the reals, to the complex. Here we examine how in the 1860's he "constructed" the real numbers from the rationals. (Check O'Connor and Robertson at St Andrews.)

(i) constructing number systems

It helps first to observe that we have already seen two instances of inventing a new system that encompasses an old one.

The first was Brahmagupta's creation of 0 and negative numbers (section IV.A.3). We said back there that you can make such things quantitative by thinking of levels ("The river is two feet below normal."). What made 0, -1, -2, … *numbers*, and not merely symbols for levels, is that Brahmagupta specified how you add and multiply among them and with the original 1, 2, …. The result was a system, the integers, that extends the set of natural numbers. It also extends the operations, meaning that the sum or product of natural numbers as integers matches their sum or product as naturals. Finally, it extends the *order*, so that the system of integers has all the properties of an ordered field except one: Not all nonzero integers have multiplicative inverses. (Go back to the pertinent axioms and check.)

The second, at the other extreme, was Bombelli's invention of a symbol i whose defining property is that you can "multiply" it by itself, with product -1. It took three centuries to find quantities describable by expressions like $a + bi$. However, Bombelli gave their arithmetic (section VI.B.4b). The result is a system that extends the (set and operations of the) field of reals and has the remarkable property that all its nonconstant polynomials have roots, but in which (*because* $x^2 + 1$ has a root) no order is possible.

(ii) Dedekind's theorem

Theorem 1. In an ordered field with the two continuum properties, suppose (only) the rational numbers

$m/n = (1 + \ldots + 1)(1 + 1 + \ldots + 1)^{-1}$ (m identities divided by n identities)

are spread between two nonempty subsets R and L. Assume that every member of R exceeds every member of L. Then the field has a member c such that for every rational q:

If $q < c$, then q is in L, and if $q > c$, then q is in R.

Recall Theorem 3 from (a) : In every ordered field, those "rationals" have to exist. To prove Dedekind's theorem, assume the LUB property.

Suppose L and R are as stated. By hypothesis, L is a nonempty set, bounded above by any of the members R is required to have. Therefore L must have an LUB; call it c.

Suppose q is rational and $q < c$. Then q cannot be an upper bound for L. Therefore there is some rational s in L exceeding q. Because q is leftward of s, q cannot be in R; it has to be in L. We have shown that if $q < c$, then q is in L.

Suppose instead $q > c$. Because c is an upper bound for L, q cannot be from L. From $q > c$, we have concluded that q is in R. That establishes the stated property of c.

The theorem is silent about where c belongs.

For an example of where c may lie, take R to be the set of rational q having

$q > 0$ and $q^2 > 2$,

L the remaining rationals. (Notice that either of R and L implies the other; we could specify just one.) Then the predicted c cannot be rational: If q is in R, then there are members of R left of q; and if q is in L, then there are members of L right of q (both Exercise 1.) Therefore c belongs to neither R nor L.

For a different example, define L_{-1} to hold the rationals Q with

$Q < 0$ and $Q^2 > 1$,

R_{-1} holding the rest. Then $c = -1$ (Exercise 2); it is the smallest member of R_{-1}.

If the second example had allowed $Q^2 \geq 1$, then c would have been the largest element of L_{-1}.

Theorem 1 assumes a partition of the rationals into left and right halves L and R. Such a partition is called a **Dedekind cut**, and the resulting c is the **cut number**, under one proviso: c is not allowed to be in L . (The reason for the rule will become clear in Exercise 3a.) We denote a Dedekind cut by $L||R$.

(iii) Dedekind's construction

The continuum properties describe the ordered field of real numbers. We have gone further, saying that they and the ordered-field axioms *characterize* the reals. "Characterize" means that any two ordered fields with the continuum properties are identical, in the "isomorphism" sense we mentioned before.

With two such fields, you can make a one-to-one correspondence $x \leftrightarrow y$ between the elements x in one and y in the other, in such a way that the additions and multiplications match:

If $x_1 \leftrightarrow y_1$ and $x_2 \leftrightarrow y_2$, then $x_1 + x_2 \leftrightarrow y_1 + y_2$ and $x_1x_2 \leftrightarrow y_1y_2$;

the orders match:

If $x_1 \leftrightarrow y_1$ and $x_2 \leftrightarrow y_2$, then $x_1 > x_2$ iff $y_1 > y_2$;

and LUB's match:

x is the LUB of the subset X iff the corresponding y is
the LUB of the corresponding subset Y.

In view of such correspondence, there exists *just one* ordered field with the Bolzano and LUB properties. However, what we know so far does not establish that such a field actually exists. Dedekind used the idea of cuts in the ordered field of rational numbers to *create* an ordered field with the properties.

Later on, we will encounter an axiomatic description of the natural numbers. That description, like our description of the reals, does not establish that a number system obeying those axioms exists. Doing so requires abstraction we will not get into. Once the system of natural numbers is in place, though, Brahmagupta's creation provides the extension (in the sense of (i)) to the integers.

Once the integers are in place, some algebraic abstraction creates an expanded system, the "rational numbers." As in the previous expansion, this one extends the set and the addition, multiplication, and order. Again, the action of the operations and order on the parent system (the integers) works as before. However, this system of rationals has the added feature of multiplicative inverses; it is a field. Upon that ordered field, Dedekind abstracted the system of "real numbers."

Dedekind *defined* a **real number** as a Dedekind cut on the rational numbers. Now, recall that a cut is a pair $L||R$ of sets that separate the rationals into left and right. How can you call those things numbers? Remember our informal rule: You get to call things "numbers" if (and only if) you specify their arithmetic. Here goes.

Defining addition is direct.

Given two cuts $L_1||R_1$ and $L_2||R_2$, by definition their sum is

$$L_1||R_1 \; + \; L_2||R_2 \qquad = \qquad (L_1 + L_2)||\{\text{all other rationals}\}.$$

The symbol $L_1 + L_2$ on the right means the rationals $l_1 + l_2$ you can make by adding l_1 from L_1 and l_2 from L_2. It is elementary to prove that $(L_1 + L_2)||\{\text{others}\}$is another Dedekind cut. (Exercise 3a gives an easy example, but still suggests how to figure the general case). Hence this addition is an operation on the set of cuts. It is clearly associative and commutative. Elementary, if lengthy, arguments establish an additive identity (Exercise 3b) and additive inverses (3c).

Accordingly, the set of cuts is an abelian group under this addition. Within that group, addition among the cuts whose cut number is rational (wherefore it is necessarily in the right half) works exactly as addition of rational numbers (Exercise 3a for evidence).

It is more complicated to define multiplication.

The intuitively obvious definition,

$$(L_1||R_1)\,(L_2||R_2) \qquad = \qquad (L_1L_2)||\{\text{rest of the rationals}\}$$

with L_1L_2 meaning the set of products l_1l_2, is not even a cut (Exercise 4a). Instead, you have to deal with separate cases.

If both L_1 and L_2 include all the negative rationals, then the definition is

$$(L_1||R_1)\,(L_2||R_2) \qquad = \qquad [L_1L_2]^+\,||\{\text{rest of the rationals}\}.$$

There, $[L_1L_2]^+$ has

(a) all the negative rational numbers and

(b) all the products l_1l_2 made from *nonnegative* l_1 and l_2 from L_1 and L_2 respectively, *if there are* any nonnegative l_1 and l_2.

(Do Exercise 4b. If in the cut $L||R$, L has all the negatives, then either L has nothing else, or L has 0. If it has only the negatives, then $L||R$ is the "zero cut," as in Exercise 3b. If it has 0, then it must also have some positives; otherwise 0 would be its biggest element.)

For the other cases, you have to use mirrors. If L_1 includes all the negatives and L_2 does not, then $-[L_2||R_2]$ (defined in Exercise 3c) does include the negatives (same exercise), the product

$$(L_1||R_1)\ (-[L_2||R_2])$$

fits the previous definition, and we define

$$(L_1||R_1)\ (L_2||R_2) \qquad = \qquad -\big((L_1||R_1)\ (-[L_2||R_2])\big).$$

We skip the evidence, as well as the details for the remaining two cases.

Once you define the multiplication, it is elementary and lengthy to prove that it is distributive and that the nonzero cuts form an abelian group under it. Consequently the cuts constitute a field.

Defining an order in that field is easy.

$(L_1||R_1) > (L_2||R_2)$ means L_1 contains (as a proper subset) L_2.

It is elementary to prove that this definition adheres to the order rules (Exercise 5 for evidence).

It is then also easy to prove as Dedekind did: For any nonempty collection of cuts, all smaller than or equal to one fixed cut, there is a least cut that is still greater than or equal to all the cuts in the collection. Thus, for any nonempty set of cuts with an upper bound, there is a least upper bound.

The construction does two things. It establishes the existence of an ordered field with the continuum properties. Since that field is unique, the construction also proves the converse of Theorem 1: In an ordered field where Dedekind cut numbers always exist, there the properties hold. The Bolzano and LUB properties are equivalent to the “Dedekind property.”

Two quotes are worth seeing here. **Struik** (page 161) says that Dedekind “accomplished for modern mathematics what Eudoxus had done for Greek mathematics” (in characterizing irrational numbers). **Struik** (160) and Boyer both translate Leopold Kronecker’s remark about the construction process: “God made the whole numbers [*ganzen Zahlen*], all others are man’s creation [*Menschenwerk*].”

Exercises IX.B.4c. In Exercises 2-5, abbreviate the subset of **Q** with some property by {property}. Thus, {< 5/3} is the set of rationals q with $q < 5/3$.

1. Show that for a positive rational q:
   a) If $q^2 > 2$, then there is a *smaller* positive rational whose square also exceeds 2.
   b) If $q^2 < 2$, then there is a bigger rational whose square is also less than 2.
   (Hint: Use the result of Exercise IX.B.4b(ii):1a.)
2. Let L_{-1} = {< 0 and with square > 1}, R_{-1} = {remaining rationals}. Show that:
   a) L_{-1} and R_{-1} define a Dedekind cut.
   b) The cut number, as defined in the theorem, is -1.
   c) That cut number is the smallest element of R_{-1}.
3. Using the definitions in this subsection, show that:
   a) $\{< 5/3\}||\{\geq 5/3\} \quad + \quad \{< -1/2\}||\{\geq -1/2\}$
   is another Dedekind cut, namely $\{< 7/6\}||\{\geq 7/6\}$.
   (You need to prove that the [apparent] sum cut accounts for all the rationals. In doing that, you will see why the definition of Dedekind cut has to specify which half the cut number, if it is rational, is allowed to be in.)
   b) $\{< 5/3\}||\{\geq 5/3\} \quad + \quad \{< 0\}||\{\geq 0\} \quad = \quad \{< 5/3\}||\{\geq 5/3\}$.
   c) Assuming that (b) establishes the additive identity,
   $-(L||R) \quad = \quad -R^*||-L^*$.
   Here, $-R^*$ means the $-r$ you can make with r from R, *excluding* the smallest r if there is one; and $-L^*$ consists of the $-l$ you make with l from L, together with $-r_0$ if r_0 is the smallest r.

4. a) Show that {< 5/3}{< -1/2}, the set of products of a rational under 5/3 with one under -1/2, cannot be the left half of a Dedekind cut.
   b) Show that the product of
   {< 5/3}||{≥ 5/3} and {< 3/4}||{≥ 3/4},
   as defined in the text, is another Dedekind cut. What cut is it?
   c) Show that by the definition in 3c,
   -({< -1/2}||{≥ -1/2}) = {< 1/2}||{≥ 1/2}.
5. Show that by the definition of order among Dedekind cuts:
   a) {< 5/3}||{≥ 5/3} > {< -1/2}||{≥ -1/2}.
   b) {< 5/3}||{≥ 5/3} + {< 3/4}||{≥ 3/4} > {< -1/2}||{≥ -1/2} + {< 3/4}||{≥ 3/4}.

5. Weierstrass

The man who recovered Bolzano's work was the most brilliant high school teacher that ever lived.

Karl (Theodor Wilhelm) Weierstrass (VYER-shtross, 1815-1897) wandered into teaching. He taught in *Gymnasia*—roughly, Prussian secondary schools—from about 1841 to 1855. As his discoveries piled up and became known, he began to receive higher offers, culminating in the university post at Berlin. Starting there at age 41, he occupied the chair for more than thirty years. His lectures became legendary for their elegance and rigor, and he maintained a prodigious output of mathematics *and mathematicians*.

The discoveries included fundamental contributions to the calculus of variations and the calculus of complex numbers. In the latter, he was the one to focus attention on the power-series representation of differentiable functions.

For us, the big interest is the precision he brought to the definitions of limit (and the related concepts of continuity, derivative, convergence) and real number. In our undergraduate calculus, the definitions are always put in his terms. More generally, in modern analysis there is "full agreement and certainty" (David Hilbert's words) about the logical validity of the foundations Weierstrass built.

a) limits and continuity

Weierstrass put symbols to Cauchy and d'Alembert's descriptions of "as little as one could wish."

Definition of limit. The real number L is the **limit** of f at $x = a$ if the following is true:
For any positive real ε (Greek lower-case *epsilon*), there is a corresponding positive δ (*delta*) such that $f(x)$ stays within ε of L as long as x, without equaling a, stays within δ of a. In symbols,

$$0 < |x - a| < \delta \qquad \text{guarantees} \qquad |f(x) - L| < \varepsilon.$$

In examples, we always gave specific ε's, like 10^{-100}. Such numerical specification is unnecessary.

We argued that the limit of $f(x) = x^2$ at $x = 2$ is 4. Let ε be a positive real.

If x is between 1 and 3, then

$$|f(x) - 4| = |x^2 - 4|$$
$$= |x + 2|\,|x - 2| \le 5|x - 2|.$$

Therefore if

$$0 < |x - 2| < \varepsilon/5, \qquad \text{then} \qquad |f(x) - 4| < 5\ \varepsilon/5 = \varepsilon.$$

We may take δ = ε/5 (or δ = 1, whichever is smaller. Why is it important that δ not exceed 1?)

Cauchy's definition of continuity was value = limit. We now put it in the symbols of Weierstrass.

(We are going to mention the set of real numbers from a to b often enough to need a symbol for it. Thus, we call the set of x with $a \le x \le b$ the **closed interval from** a **to** b, and abbreviate it as $[a, b]$.)

Definition of continuity. Suppose *f* is defined on (meaning at each place in) the closed interval [*a*, *b*]. Given *c* in that interval, we say *f* **is continuous** at $x = c$ if for any positive real ε, there is a corresponding positive δ such that

$$|x - c| < \delta \qquad \text{guarantees} \qquad |f(x) - f(c)| < \varepsilon$$

(it being understood that *x* is also in the interval).

The Weierstrass notation would be an advance if all it did was give us generality, a way to avoid giving examples like $\varepsilon = 10^{-6}$. However, it also led to an important concept in continuity that Bolzano had foreshadowed.

> Look at $f(x) = x^2$ on the interval [1, 3]. If *c* and *x* are in the interval, then
> $$|f(x) - f(c)| = |x + c|\,|x - c| \leq 6\,|x - c|.$$
> Therefore given $\varepsilon > 0$, we can force
> $$|f(x) - f(c)| < \varepsilon$$
> by keeping
> $$|x - c| < \varepsilon/6.$$

Certainly, that says *f* is continuous at every *c* in [1, 3]. But the continuity has a special character. If you name a positive ε, then there is a corresponding $\delta = \varepsilon/6$ *that works everywhere in the interval*. In that situation, we say *f* is **uniformly continuous** on the interval. That is a distinction Cauchy's verbal description did not make.

> Look at
> $$g(x) = 1/x \qquad \text{(strictly) between } x = 0 \text{ and } x = 1.$$
> It is continuous at $x = 0.9$. Take any $\varepsilon > 0$. If we restrict *x* to $0.8 < x < 1.0$, then
> $$\begin{aligned}|g(x) - g(0.9)| &= |1/x - 1/0.9| \\ &= |0.9 - x|/(.9x) \leq |0.9 - x|/(.8)^2.\end{aligned}$$
> Therefore if *x* also satisfies
> $$|0.9 - x| < \delta_1 = 0.64\varepsilon,$$
> then necessarily
> $$|g(x) - g(0.9)| < \varepsilon.$$
> It is also continuous at $x = 0.009$. Take that same $\varepsilon > 0$. If we restrict *x* to $0.008 < x < 0.010$, then
> $$\begin{aligned}|g(x) - g(0.009)| &= |0.009 - x|/(.009x) \\ &\leq |0.009 - x|/(.008)^2.\end{aligned}$$
> To guarantee that
> $$|g(x) - g(0.009)| < \varepsilon,$$
> now we have to squeeze *x* into
> $$|0.009 - x| < \delta_2 = 0.000064\,\varepsilon.$$

Our *g* is continuous everywhere between $x = 0$ and $x = 1$, but keeping $g(x)$ close to $g(c)$ requires tighter restriction on *x* as $c \to 0$. On this interval, *g* is not uniformly continuous. (Compare Exercise 2.)

You can supply for yourself the epsilon-delta definition of differentiability and derivative. The new element Weierstrass brought to them was an example of a continuous function that is not differentiable *anywhere* (see Wikipedia®). Recall the intuitive notion of continuous function: one whose graph you can draw without picking the pencil off the paper. (Recall also Dirichlet's example of a function not continuous anywhere). Even after Cauchy's definition of continuity, it was thought that a continuous function's graph would be mostly smooth, having a non-vertical tangent at all but some separated points (as the graph of the cycloid has). The Weierstrass example destroys that notion.

Oddly, or maybe appropriately, the Weierstrass example (1872) did not come to light until after Bernhard Riemann had published one, and itself came after another unknown example from Bolzano.

Exercises IX.B.5a

1. Prove that if f and g are continuous at $x = c$, then so is the sum $f + g$.
2. a) Show that $f(x) = x^2$, as defined for all x, is not *uniformly* continuous.
   b) Show that $g(x) = 1/x$, defined for $1 \leq x \leq 2$, *is* uniformly continuous.

b) maxima and minima

Weierstrass derived a property of continuous functions that advanced the studies of both the functions and the nature of the real numbers.

Theorem 1. (The Extreme-Value theorem) Suppose f is continuous on the closed interval $[a, b]$. Then there are points c and d in the interval such that for every x there,

$$f(d) \leq f(x) \leq f(c).$$

In words, f has a smallest value and a biggest value in the interval.

To see that the theorem is news, let exceptions prove (meaning "test") the rule.

First, the conclusion may fail if either endpoint is missing. Return to

$$g(x) = 1/x \qquad 0 < x < 1.$$

If you name any a in the interval, then there are points x to its left (like $x = a/2$) where $g(x) > g(a)$ and others to its right (Name one.) where $g(x) < g(a)$. Therefore g has neither biggest nor smallest value. We can provide for the latter by allowing $0 < x \leq 1$, which would make $g(1)$ the minimum. But with $\lim_{x\to 0} g(x) = \infty$, no maximum is possible without removing an *interval* rightward from 0.

Second, the conclusion may fail if the function is not continuous. Recall (section IX.B.2c) how we said that the Fourier series

$$F(x) = \pi - 2/1 \sin x - 2/2 \sin 2x - \ldots$$

has values

$$F(x) \quad = x \qquad \text{if } 0 < x < 2\pi$$
$$= \pi \qquad \text{if instead } x = 0 \text{ or } x = 2\pi.$$

From the $F(x) = x$ part, we can see that

$$\lim_{x\to 0} F(x) = 0 \neq F(0), \qquad \lim_{x\to 2\pi} F(x) = 2\pi \neq F(2\pi).$$

The function is not continuous, and it does not have extreme values. On the interval $[0, 2\pi]$, there are places where $F(x) \approx 2\pi$. But F never reaches the value 2π; no value of F is a maximum. Similarly, F lacks a minimum value. (Compare Exercise 1.)

To prove Theorem 1, we first need a characteristic of the reals. It is one of the results of Bolzano that Weierstrass rediscovered.

Theorem 2. (The Bolzano-Weierstrass Theorem) If an infinite set of real numbers is bounded, then it has an accumulation point.

The set S is **bounded** if it is bounded below and above: There exist reals m and M such that every member of S is between them. The real number r is an **accumulation point** of S if, for any positive ε, the interval from $r - \varepsilon$ to $r + \varepsilon$ includes an infinity of members of S.

Abbreviate "accumulation point" by "AC." An AC of S may or may not be in S. The set

$$S = \{1, 1 + 1/2, 1/4, 1 + 1/8, 1/16, 1 + 1/32, 1/64, \ldots\}$$

has AC's 0 and 1 (Exercise 2a) Of those two, 1 in in S, 0 is not. Note that 1/4 is in S, but is not an AC (Exercise 2b).

For an exception, observe that the natural numbers form an unbounded infinite set without AC's (Exercise 2c). At the other extreme, *every real number* is an AC for the set of rationals. (Reason?)

We will derive Theorem 2 from the LUB property. It happens that the theorem is *equivalent* to the continuum properties: In an ordered field where Theorem 2 is true, the LUB property holds. After you see the proof below, try to prove the equivalence in Exercise 3.

To save some symbols, assume the infinite set S is bounded by 0 and 1.

Look at the two intervals [0, 1/2] and [1/2, 1]. At least of them must have infinitely many members of S. Call it $[a_1, b_1]$, and notice that

$$0 \le a_1 < a_1 + 1/2 = b_1 \le 1.$$

One of the intervals $[a_1, (a_1 + b_1)/2]$, $[(a_1 + b_1)/2, b_1]$ must have infinitely many member of S. Call it $[a_2, b_2]$, and notice that

$$a_1 \le a_2 < a_2 + 1/4 = b_2 \le b_1.$$

Continuing that way indefinitely, we end up with two sequences having

$$a_1 \le a_2 \le \ldots, \qquad b_1 \ge b_2 \ge \ldots, \qquad \text{and} \qquad b_n = a_n + 1/2^n.$$

The (possibly finite) set of numbers

$$T = \{a_1, a_2, \ldots\}$$

is nonempty and bounded. Let c be its LUB. If ε is any positive real, then $c - \varepsilon$ is not an upper bound for T. That means some a_k exceeds $c - \varepsilon$. That implies in turn that

$$a_{k+1}, a_{k+2}, \ldots,$$

being at least a_k and not more than c, are likewise right of $c - \varepsilon$. As soon as $n \ge k$ is big enough to make $1/2^n < \varepsilon$, we have

$$c - \varepsilon < a_n \le c \qquad \text{and} \qquad b_n = a_n + 1/2^n < c + \varepsilon.$$

Since the interval $[a_n, b_n]$ has an infinity of members of S, those members are all between $c - \varepsilon$ and $c + \varepsilon$. That proves c is an accumulation point of S.

With Theorem 2, we can prove an important property of continuous functions.

Theorem 3. A function continuous on a closed interval is bounded there.

We say a function is **bounded** if its values are all between some fixed reals. From the example of

$$g(x) = 1/x, \qquad 0 < x < 1,$$

we already know that not all continuous functions are bounded.

To prove Theorem 3, suppose that the values of f on $[a, b]$ are not bounded above. Then 1 is not an upper bound, so there exists in the interval a place x_1 where $f(x_1) > 1$. For the same reason there is a place x_2 where

$$f(x_2) \quad \text{exceeds} \quad \text{both 2 and } f(x_1),$$

a place x_3 where

$$f(x_3) \quad \text{exceeds} \quad \text{all three of 3, } f(x_1), \text{ and } f(x_2),$$

and so on. By the method of selection, all of $x_1, x_2, \ldots$ are different numbers in $[a, b]$. Therefore the set $S = \{x_1, x_2, \ldots\}$is infinite. The Bolzano-Weierstrass theorem guarantees an accumulation point r.

This r has to be in the interval. It cannot be, for example, $d = b + 0.001$. The gap from $d - 0.0005$ to $d + 0.0005$ lies entirely outside $[a, b]$, is therefore devoid of x_n's. That means d is not an accumulation point of S, and $r \ne d$.

At $x = r$, f is not continuous. If it were, then by the definition there would be a $\delta > 0$ such that

$$|x - r| < \delta \qquad \text{guarantees} \qquad |f(x) - f(r)| < 1.$$

In words, for every x between $r - \delta$ and $r + \delta$, $f(x)$ would be between $f(r) - 1$ and $f(r) + 1$. But there is an infinity of x_n's between $r - \delta$ and $r + \delta$; and at each one with index n exceeding $f(r) + 1$,

$$f(x_n) \;>\; n \;>\; f(r) + 1.$$

We see that if f is not bounded above, then it is not continuous someplace in $[a, b]$. The same holds if f is not bounded below. The contrapositive is Theorem 3.

Now we have what we need to prove the extreme-value theorem.

Assume f is continuous on $[a, b]$. The set of values $f(x)$ is certainly nonempty, and Theorem 3 tells us it is bounded. Therefore it has a least upper bound; call the LUB M. We are going to find a point c where $f(c) = M$. That will spot the biggest value of f, and similar reasoning finds the smallest.

Because $M - 1$ is not an upper bound, there must be a place x_1 where

$$f(x_1) > M - 1.$$

It might be that $f(x_1) = M$. Then we take $c = x_1$. The alternative is $f(x_1) < M$. In that case, neither $f(x_1)$ nor $M - 1/2$ is an upper bound, so there must be a place x_2 where

$$f(x_2) > f(x_1) \qquad \textit{and} \qquad f(x_2) > M - 1/2.$$

Next, either $f(x_2) = M$ and we take $c = x_2$, or instead $f(x_2) < M$ and we keep going. You see the pattern: We continue until some $f(x_n) = M$, at which time we take $c = x_n$ and stop; or else we establish an unending sequence $x_1, x_2, \ldots$ that has

$$f(x_1) \;<\; f(x_2) \;<\; f(x_3) \;<\ldots \qquad \text{and} \qquad \text{every } f(x_n) > M - 1/n.$$

If the process is unending, the numbers $x_1, x_2, \ldots$ are all unequal, because they give different values of f. Hence they constitute an infinite set. By the Bolzano-Weierstrass theorem, the set $\{x_1, x_2, \ldots\}$ has an accumulation point c. By our argument for Theorem 3, c has to be in $[a, b]$. Because M is the LUB of the values, we must have

$$f(c) \le M.$$

It cannot be that $f(c) < M$. Imagine that $f(c)$ were 1.999 and M were 2. Because f is continuous at $x = c$, there would exist a positive δ such that

$$c - \delta \;<\; x \;< c + \delta \qquad \text{forces} \qquad 1.9985 \;<\; f(x) \;<\; 1.9995.$$

Because c is an accumulation point of $\{x_1, x_2, \ldots\}$, there must be an infinity of x_n between $c - \delta$ and $c + \delta$. Infinitely many of those have $n > 10{,}000$. For any such x_n,

$$f(x_n) \;>\; M - 1/n \;>\; 2 - 0.0001.$$

We would have simultaneously

$$f(x_n) > 1.9999 \qquad \text{and} \qquad f(x_n) < 1.9995,$$

a contradiction. It must be that $f(c) = M$.

Exercises IX.B.5b

1. We know $f(x) = x^2 - 3x$ is continuous on $[0, 5]$. What are its biggest and smallest values in the interval? (The answer is doable by just algebra.)
2. Show that:
   a) 0 and 1 are accumulation points of
   $S = \{1, 1 + 1/2, 1/4, 1 + 1/8, 1/16, 1 + 1/32, 1/64, \ldots\}$.
   b) 1/4 is not an accumulation point of S.
   c) $\{1, 2, 3, \ldots\}$ has no accumulation points.
3. In an ordered field, assume that Theorem 2 is true. Prove the LUB property.
   (The problem is elementary, relative to the level we have reached, but hard. Hints: If S is bounded above, then either it has a biggest member, or it does not; and if s is in S and M is an upper bound, then $(s + M)/2$ is an upper bound, or is not.)

c) arithmetization of the reals

We had barely thought about sequences, and suddenly they popped up in all the arguments of (b) above. Weierstrass showed how to recast the definitions from calculus—function limit, continuity, derivative—into the language of sequences. Rather than doing the recasting, we will look at sequences themselves and some of their properties. That will allow us to show how Weierstrass *defined* the real numbers in terms of sequences.

(i) convergence of sequences

We will allow ourselves the intuitive notion of a sequence

$a_1, a_2, a_3, \ldots$

(which we will mostly abbreviate by (a_n)) as a parade of real numbers. [The technical definition is that a **sequence** is a function A that assigns the real number $A(n)$ to the natural number n. With sequences, it is usual to write a_n instead of $A(n)$.] The difference between the sequence and the *set* $\{a_1, a_2, a_3, \ldots\}$is that in the sequence, order counts. Thus,

$1, 2, 1, 2, 1, 2, \ldots$ and $1, 1, 2, 2, 1, 1, 2, 2, \ldots$

are unequal sequences, even though the collections of values are the same.

We say the sequence (r_n) **converges to** (or **tends to**, **approaches**, or **has as limit**) the real number L if corresponding to any positive ε, there is a natural N such that

$n > N$ guarantees $|r_n - L| < \varepsilon$.

We met the definition before, in the specific setting of the sequence (s_n) of partial sums of a *series* (section IX.B.3b). Match the definition back there, which used Cauchy's words, against the one here, which uses Weierstrass's symbols.

The most elementary theorem about sequences is for monotonic sequences. We say a sequence (r_n) is **increasing** if

$r_1 \le r_2 \le r_3 \le \ldots.$

We say (s_n) is **decreasing** if

$s_1 \ge s_2 \ge s_3 \ge \ldots.$

Either kind is called **monotonic**. [With ≤ instead of <, the correct mathematical term is **nondecreasing**, as is **nonincreasing** for ≥. Since the usual mathematical choice is inclusiveness—as in the usage of "or"—we'll stay with "increasing" and "decreasing."]

Theorem 1. If a sequence is monotonic *and bounded*, then it converges to a real number

The proof is Exercise 1, but the needed argument is actually buried within our proof of the Bolzano-Weierstrass theorem (section IX.B.5b).

Theorem 1 applies directly to *series* of nonnegative terms. If each $a_k \ge 0$, then the partial sums

$s_n = a_1 + a_2 + \ldots + a_n$

form an increasing sequence. If the sequence is bounded, then the series $a_1 + a_2 + \ldots$ converges to a real number. If instead (s_n) is unbounded, then the series converges to infinity. See Exercise 2.

(ii) Bolzano's criterion

One of the most fundamental ideas in all of analysis identifies the sequences that converge.

Theorem 2. (Bolzano's Criterion) The sequence (r_n) converges (to a real limit) exactly if it has the property that for each $\varepsilon > 0$, there is a corresponding N past which the terms are within ε:

If both m and n exceed N, then $|r_m - r_n| < \varepsilon$.

Look back at "Cauchy's criterion" (for series in section IX.B.3b). It was Bolzano's idea first, rendered here with the symbols of Weierstrass. A sequence that meets the criterion is called a **Cauchy**

sequence. [Yes, it pains me not to say “Bolzano sequence.” Unfortunately, “Cauchy sequence” is universal, and the concept is enormously important. I have to yield.] Earlier, we argued by example why a *function* possessed of a limit must meet the corresponding (function) criterion. Do Exercise 3 for sequences. We will tackle the tougher part, proving that a Cauchy sequence must converge.

To prove Theorem 2, assume that (r_n) satisfies the criterion.

First, we establish that the sequence has to be bounded. Name the specific $\varepsilon = 1$. There must exist N beyond which

m and n both exceed N forces $|r_m - r_n| < 1$.

Write

$K = |r_1| + |r_2| + \dots + |r_N| + |r_{N+1}| + 1.$

Clearly each of

$|r_1|, \quad |r_2|, \quad \dots \quad |r_N|, \quad |r_{N+1}|$

is smaller than K. The remaining

$|r_{N+2}|, |r_{N+3}|, |r_{N+4}|, \dots$

all have

$|r_m - r_{N+1}| < 1.$

That puts each such r_m strictly between $[r_{N+1} - 1]$ and $[r_{N+1} + 1]$, so that

$|r_m| < |r_{N+1}| + 1 \le K.$

Hence K is an upper bound, $-K$ a lower bound. The sequence is bounded.

Second, we show that there must be a real L such that for any $\varepsilon > 0$, the interval between $L - \varepsilon$ and $L + \varepsilon$ has an infinity of terms from (r_n). (Be careful: You must remember that an infinity of *terms* might not constitute an infinite set of real values.) It could be that the set

$S = \{r_1, r_2, r_3, \dots\}$

is finite. There is only one way that could happen: At least one value is repeated infinitely many times. Imagine that

$r_2 = r_4 = r_8 = r_{16} = \dots.$

In that case, we simply take $L = r_2$. The alternative is that S is infinite. Then S is a bounded infinite set. By the Bolzano-Weierstrass theorem, S has an accumulation point L. By definition, infinitely many members of S—and therefore an infinity of terms of the sequence—lie between $L - \varepsilon$ and $L + \varepsilon$. That proves the second contention. (Along these lines, do Exercise 4.)

Third and last, we prove that the sequence converges to L. If ε is positive, then so is $\varepsilon/2$. Because the sequence is Cauchy, there exists M such that

m and n both exceed M forces $|r_m - r_n| < \varepsilon/2$.

Among $r_{M+1}, r_{M+2}, \dots$, infinitely many must be between $L - \varepsilon/2$ and $L + \varepsilon/2$. Let r_N be any one of them. Then for all $n > N$,

r_n is within $\varepsilon/2$ of r_N and r_N is within $\varepsilon/2$ of L.

That puts r_n within $\varepsilon/2 + \varepsilon/2$ of L. We have proved that L is the limit of (r_n).

Exercises IX.B.5c

1. Prove that if (r_n) is increasing and bounded, meaning

   $r_1 \le r_2 \le r_3 \le \dots$ and every $r_n <$ some fixed M,

   then the sequence converges to a real number. (Hint: LUB property.)

2. Suppose $a_1 + a_2 + \dots$ is a series of nonnegative terms.

   a) Show that if the sequence of partial sums

   $s_n = a_1 + a_2 + \dots + a_n$

is bounded, then the series converges.
b) Show that if (s_n) is not bounded, then the series meets the definition (answering Exercise IX.B.3b:2) for convergence to infinity.
c) Does either (a) or (b) stay true if some of the a_k are negative?

3. Prove that if (r_n) converges to a real number, then for any $\varepsilon > 0$, there exists N such that
   m and *n* both exceed *N* forces $|r_m - r_n| < \varepsilon$.
4. Show that if a sequence is bounded, then some subsequence of it converges to a real number. (A **subsequence** of (r_n) is a sequence assembled by choosing terms from (r_n) in correct order [in sequence?], as with
   $r_2, r_4, r_8, r_{16}, \ldots.$
   The statement here is also called “the Bolzano-Weierstrass theorem.” Try our methods, then look at the marvelous proof (using just Theorem 1) in Kenneth A. Ross’s *Elementary Analysis: The Theory of Calculus* (page 53, referring back to pages 51-52, in the 1980 edition, published by Springer).)
5. Show that if (w_n) and (x_n) are Cauchy, then so are ($w_n + x_n$) and ($w_n x_n$). (For the latter, it turns out to be important to recall that a Cauchy sequence is necessarily bounded.)

(iii) the construction of the reals

Recall Dedekind’s construction of an ordered field with the LUB property (section IX.B.4c(iii)). He interpreted pairs of sets partitioning the rational numbers, “Dedekind cuts,” as numbers. Weierstrass (1860’s) interpreted as numbers sets of Cauchy sequences. (That is a somewhat higher abstraction. A single Dedekind cut represents a real number. It takes an infinite set of sequences to do the same.)

Let us, like Dedekind, begin by looking at just the ordered field of rational numbers. View the sequences

(q_n): 1+1/2, 1+1/4, 1+1/8, 1+1/16, …, (r_n): 1/2, 2/3, 3/4, 4/5, ….

They are unequal, but have the same rational limit, 1. Because they converge, they have to be Cauchy. We will call them “equivalent,” and they and the other sequences of rationals converging to 1 form their “(equivalence) class.” For Weierstrass, that class *was* the real number 1.

Look next at the two sequences

(s_n): 2, 7/4, 97/56, 18817/10864, …, (t_n): 1, 1.7, 1.73, 1.732, ….

The (s_n) sequence comes from the Babylonian square-root algorithm, reflected for √3 in section III.A.8a. It is the algorithm’s nature to produce decreasing rational overestimates whose separation from the next estimate approaches half the separation from the previous:

$$0 < s_n - s_{n+1} \approx 1/2\,(s_{n-1} - s_n).$$

For that reason, (s_n) is Cauchy. The (t_n) sequence is from the Indian (decimal) square-root algorithm (section IV.A.2). Because the algorithm produces immutable digits, if $m > n$, then

$$0 \le t_m - t_n \le 10^{-(n-1)};$$

(t_n) is also Cauchy. Judging from

$$s_1^2 = 4, \quad s_2^2 = 3.0625, \quad s_3^2 \approx 3.0003, \quad s_4^2 \approx 3.000\,000\,008,$$
$$t_1^2 = 1, \quad t_2^2 = 2.89, \quad t_3^2 \approx 2.9929, \quad t_4^2 \approx 2.999\,8,$$

each sequence approaches a number whose square is 3. There is no such number; our field of vision, remember, is limited to **Q**. However, from

$$s_n^2 \approx 3 \approx t_n^2,$$

we conclude

$$s_n - t_n = (s_n^2 - t_n^2)/(s_n + t_n) \approx 0/3.5.$$

The full definition is that (u_n) is **equivalent to** (v_n) if

$$\lim_{n\to\infty}(u_n - v_n) = 0.$$

In the example above, (s_n) and (t_n) are equivalent. The set of Cauchy sequences equivalent to both is called their **equivalence class**. In the Weierstrass construction, this class is the real number √3.

In that construction, the operations and the order are easy to define. Assume (w_n) and (x_n) are Cauchy sequences of rational numbers. Then the operations are defined by

[class of (w_n)] + [class of (x_n)] = [class of $(w_n + x_n)$],
[class of (w_n)] × [class of (x_n)] = [class of $(w_n x_n)$].

It is easy to prove that $(w_n + x_n)$ and $(w_n x_n)$ are also Cauchy sequences (Exercise 5 above). That is the end of the easy part. The hard part, requiring work with abstraction, is to show that these "+" and "×" really do constitute operations on the collection of *classes* of Cauchy sequences.

> For an example of what is required, recall that (in our examples)
>
> [class of (q_n)] and [class of (r_n)]
>
> are the same class. You would have to prove that "adding" either of them to [class of (s_n)] gives the same result:
>
> [class of (q_n)] + [class of (s_n)] and [class of (r_n)] + [class of (s_n)]
>
> are the same class. Part of the difficulty is that you have to become familiar with equivalence relations (see Exercise VIII.C.2a:3) and equivalence classes. We will skip the work, and merely state that under these "operations," the set of equivalence classes constitutes a field.

Next, define the order by defining

[class of (w_n)] > [class of (x_n)]

to mean the w_n **eventually** exceed the x_n by at least a fixed margin. In symbols, there is a positive rational c and some natural N such that

for every $n > N$, $w_n > x_n + c$.

That ends the order's easy part. It turns out to be a project to show that the definition is valid, that it turns the set of classes into an ordered field, and that the field has the LUB property. Thus did Weierstrass put together his model of the real number system.

> You might be wondering about "eventually" and that "margin" c. It is not necessary for *all* the terms of (w_n) to exceed those of (x_n), and it is not sufficient.
>
> In the two Cauchy sequences
>
> (a_n): $10^{100}/1, 10^{100}/2, 10^{100}/3, \ldots$ and (b_n): $1 - 10^{100}/1, 1 - 10^{100}/2, 1 - 10^{100}/3, \ldots$,
>
> the first $(2\times10^{100} - 1)$ terms of (a_n) exceed those of (b_n). But eventually, namely beyond $N = 3\times10^{100}$,
>
> $a_n < 1/3$ and $b_n > 2/3 > a_n + 1/3$.
>
> In our earlier examples,
>
> every $q_n = 1 + 1/2^n$ exceeds every $r_n = n/(n + 1)$.
>
> However,
>
> [class of (q_n)] > [class of (r_n)]
>
> is not permissible; those two are the same class. The definition above,
>
> [class of (w_n)] > [class of (x_n)] provided eventually $w_n > x_n + c$,
>
> guarantees that the limit of (w_n)—a rational number or else a hole in the rationals—must *exceed* the limit of (x_n).

6. Riemann

Struik (page 156) describes Georg Friedrich Bernhard Riemann (REE-mahn, 1826-1866) as "the man who more than any other has influenced the course of modern mathematics." That is strong praise for a man who did not live to forty. Riemann's definition of integral, which is our interest, led to discoveries that made the theory of integration a whole new area of analysis. His work on surfaces, extending discoveries of Gauss, led to a whole new kind of geometry. That study is still an important area of research, and is essential to the theory of relativity. Part of his study of surfaces was based on his discoveries about functions of a complex variable, in which he elaborated the fundamental relations called the "Cauchy-Riemann equations." Those are PDEs, influenced by Riemann's interest in hydrodynamics, which was just one area of mathematical physics to which he contributed.

Riemann submitted his doctoral dissertation in 1851, to Gauss. When Dirichlet died in 1859, Riemann succeeded him in Gauss's old chair. Aside from the study of surfaces, Riemann pursued results of Gauss in number theory, specifically the prime number theorem (section VIII.C.2c). In that connection, Euler had looked at the function F defined for real $x > 1$ by the series

$$F(x) \;=\; 1/1^x + 1/2^x + 1/3^x + \ldots.$$

[Recall (section IX.B.1) that Euler had evaluated $F(2)$. You see the levels where Riemann operated: Euler, Gauss, Cauchy, Dirichlet.] Riemann replaced x by the complex variable z, showed how the definition of F could be extended to all complex z, and studied the resulting "zeta function" ζ (Greek letter zeta). He conjectured that if

$$z = a + bi \text{ is not real } (b \neq 0) \qquad\qquad \text{and} \qquad\qquad \zeta(z) = 0,$$

then necessarily $a = 1/2$. With Fermat's last theorem settled, that "Riemann hypothesis" is the most famous unsolved problem in mathematics.

Riemann's definition, around 1850, completed rigorous characterization of integrals, just as Weierstrass had rigorously characterized continuity and differentiability.

a) integrability and integrals

(i) extreme values and limits

Suppose f is defined on a closed interval. Recall Fermat (section VII.A.4e) and Leibniz (section VII.B.2): They broke up the interval into subintervals; formed sums of the form

(value #1 of f)[width of subinterval #1] + (value #2 of f)[width of subinterval #2] + …

(albeit, for Fermat, with an infinity of terms); and in effect found the limits of the sums as the widths dropped toward zero. The convenient values of f were the maximum and minimum within the subintervals. For friendly functions, those extremes are easy to spot, and they help determine the limits.

(ii) bounds instead of extremes

By the time of Riemann, two difficulties were glaring. First is that a function might not have extremes in a subinterval. Our favorite functions are monotonic—they increase, or they decrease—whose extremes are at the ends. If a function is at least continuous, then we know its extremes exist. (Reason?) But recall that a discontinuous function may fail to have extremes, even if it has bounds.

> To illustrate the difficulty, modify Dirichlet's function D (section IX.B.3a(ii)) on [0, 1] to write
>
> $$d(x) \;=\; xD(x) \;=\; \begin{array}{ll} x & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational.} \end{array}$$
>
> Clearly d is bounded: $0 \;\le\; d(x) \;\le\; 1$.
>
> The function is discontinuous everywhere *except at* $x = 0$. Near the rational $a = 1/2$, between $x = a - 10^{-9}$ and $x = a + 10^{-9}$, there are rational s with $d(s)$ within 10^{-9} of $d(a)$, but also irrational t with $d(t)$ fully 1/2 from $d(a)$. The mirror image happens near the irrational $b = \sqrt{2}/2$. Only at $c = 0$ is it true that $d(x) \approx d(c)$ for all nearby x, so that d is continuous.

On the subinterval $[a, b]$, d has a smallest value but no largest. At any irrational t there, $d(t) = 0$, reaching the minimum. At any rational s, $d(s) < \sqrt{2}/2$, and just left of $x = b$ there are rationals r where $d(r) \approx \sqrt{2}/2$. Thus, $\sqrt{2}/2$ is the least upper bound, but not itself a value, of d on the subinterval.

See also in section IX.B.5b the description of the values over $[0, 2\pi]$ of the Fourier series

$$F(x) = \pi - 2/1 \sin x - 2/2 \sin 2x - \ldots.$$

Riemann finessed the extreme-value difficulty by defining upper and lower estimates of integrals in terms of least upper bounds and greatest lower bounds.

Assume now that f is bounded on its interval of definition, say

$$m \le f(x) \le M \qquad \text{for all } x \text{ in } [a, b].$$

Subdivide (hereafter "partition") the interval into k subintervals (which might have unequal widths). On subinterval #i, whose width is w_i, symbolize the least upper bound of f by L_i and the greatest lower bound by G_i. Form the partition's **upper sum**

$$u = L_1 w_1 + L_2 w_2 + \ldots + L_k w_k$$

and **lower sum**

$$l = G_1 w_1 + G_2 w_2 + \ldots + G_k w_k .$$

Those correspond to our old upper and lower estimates for the integral.

(iii) bounds instead of limits

Having such estimates, Fermat and Leibniz had proceeded to their common limits. The second difficulty lies there. We now know that variable quantities may fail to have limits.

However, our sums definitely have bounds: Both l and u lie between

$$mw_1 + mw_2 + \ldots + mw_k = m[b - a] \quad \text{and} \quad Mw_1 + Mw_2 + \ldots + Mw_k = M[b - a].$$

Therefore they have least upper and greatest lower bounds. Riemann called the GLB of the possible upper sums, considering all possible partitions of $[a, b]$, the **upper integral** of f. Similarly he called the LUB of the lower sums the **lower integral**. If those match, then the function **is integrable on** $[a, b]$, and their common value is by definition the **integral**.

b) three examples

To see the significance of the definition, it helps to see a function that does not fit it.

Example 1. Return to the function on [0, 1] given by

$$d(x) = \begin{array}{ll} x & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational.} \end{array}$$

Partition the interval into the three subintervals $[0, 1/2]$, $[1/2, \sqrt{2}/2]$, $[\sqrt{2}/2, 1]$.

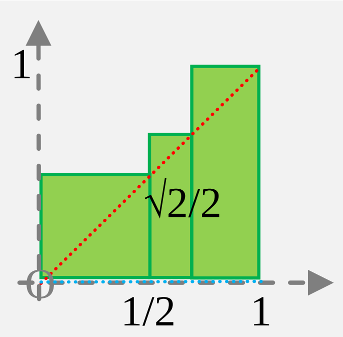
In each subinterval, d reaches a minimum value 0 at any irrational. Therefore the lower sum is

$$0\,[1/2 - 0] + 0\,[\sqrt{2}/2 - 1/2] + 0\,[1 - \sqrt{2}/2] = 0.$$

That happens no matter how you slice it—no matter what the partition is. Therefore the lower integral, the LUB of the lower sums, is 0.

In the first and last subintervals, d reaches a maximum at the right endpoint. In $[1/2, \sqrt{2}/2]$, there is no maximum, but the LUB of d is $\sqrt{2}/2$. Therefore the upper sum is

$$1/2\,[1/2 - 0] + \sqrt{2}/2\,[\sqrt{2}/2 - 1/2] + 1\,[1 - \sqrt{2}/2].$$



View the figure at right. It tries to suggest the graph of d with one dotted red part for rational x, one horizontal dotted blue for irrational. The upper sum, adding the areas of the green rectangles, exceeds the area of the triangle with vertices (0, 0), (0, 1), and (1, 1). Regardless of the partition, the upper sum

(right-end value #1)[width #1] + … + (right-end value #*k*)[width #*k*]
exceeds 1/2. Therefore the upper integral is at least 1/2. (Do you see why it is *exactly* 1/2?) The upper integral does not match the lower. The function is not integrable.

As you might expect, our old friends do not misbehave that way.

Example 2. Look (as in section IX.B.5a) at $f(x) = x^2$ on the interval [1, 3].

Partition the interval into $k = 10^9$ equally wide subintervals. Call their endpoints

$x_0 = 1,\qquad x_1 = 1 + 2\times10^{-9},\qquad x_2 = 1 + 4\times10^{-9},\qquad \ldots,\qquad x_k = 1 + 2k\times10^{-9}.$

In subinterval #*j*, *f* reaches its maximum x_j^2 at the right endpoint and its minimum x_{j-1}^2 at the left. Those values differ by

$$\begin{aligned} x_j^2 - x_{j-1}^2 &= (x_j - x_{j-1})\,(x_j + x_{j-1}) \\ &= 10^{-9}\,(x_j + x_{j-1}) \\ &< 10^{-9}\,(3 + 3). \end{aligned}$$

Therefore the upper sum

$$u = x_1^2\,[10^{-9}] + x_2^2\,[10^{-9}] + \ldots + x_k^2\,[10^{-9}]$$

and the lower sum

$$l = x_0^2\,[10^{-9}] + x_1^2\,[10^{-9}] + \ldots + x_{k-1}^2\,[10^{-9}]$$

differ by

$$\begin{aligned} u - l &= (x_1^2 - x_0^2)[10^{-9}] + \ldots + (x_k^2 - x_{k-1}^2)[10^{-9}] \\ &< 10^{-9}\,(3 + 3)\,[1]. \end{aligned}$$

Take that paragraph as evidence that you can find upper sums as close as you want to lower sums. Therefore the GLB of the upper sums is less than or equal to the LUB of the lowers:

upper integral $\leq$ lower integral.

[Next take my word for it that the upper integral cannot be less than the lower.] It follows that upper and lower integral are equal; *f* is integrable on its interval.

Compare the short treatment of the same function in section IX.B.5a. You should see that the key to the middle paragraph in Example 2 is that $f(x) = x^2$ is *uniformly continuous* on [1, 3]. The same argument will show that any uniformly continuous function is integrable on its interval.

In 1872, Heinrich Eduard Heine (1821-1881) published the statement and proof that if a function is continuous on a closed interval, then it is uniformly continuous there. That result showed that Riemann's characterization of integrable functions encompasses all continuous functions.

[Bolzano knew the statement by the 1830's, and Dirichlet produced proof in 1854.

Heine, who studied under Weierstrass, is sometimes called by his middle name, to distinguish him from the poet Heinrich Heine. Following Dirichlet's method, Eduard established a certain property of closed intervals. Émile Borel (1871-1956) abstracted that property. In analysis, the Heine-Borel property proved to be a powerful weapon. In topology, it became one of the most fundamental concepts.]

Integrability of continuous functions might lead us to believe that continuity is essential. Riemann produced a remarkable example to show that the connection is not so easy.

Example 3. Riemann defined the function given by

$$R(x) = \begin{cases} 0 & \text{if } x \text{ is irrational} \\ 1/n & \text{if } x \text{ is the reduced fraction } m/n. \end{cases}$$

Look at it on the interval [3, 4] (because we want to refer to π).

First, the function has the odd property of being discontinuous at the rationals *but continuous at all irrationals*. It is discontinuous at $a = 3.5 = 7/2$, because $R(a) = 1/2$ whereas $R(x) = 0$ at the nearby irrationals. It is continuous at π, because $R(\pi) = 0$ and at all nearby *x*'s $R(x)$ is *small*.

To see that last part, hark back to Exercise III.A.6b:4. It asked for fractions closer to π than 22/7 is. A spreadsheet or programmable calculator can determine this for you: Of the fractions between 3 and 4 whose denominators are 100 or less, the closest to π is 3 + 14/99 = 311/99. Now

$$\pi - 311/99 \approx 3.141\,593 - 3.141\,414 > 0.000\,178.$$

Therefore if x is between $\pi - 0.000\,17$ and $\pi + 0.000\,17$, then either x is irrational and $R(x) = 0$, or else x is a rational number whose reduced denominator exceeds 100 and

$$0 < R(x) - R(\pi) < 1/100.$$

Analogously we can show that $R(x)$ can be put within any chosen ε of $R(\pi)$; R is continuous at $x = \pi$.

Second, this highly discontinuous function is integrable. Consider the list of fractions

3 + 1/2
3 + 1/3,   3 + 2/3
3 + 1/4,   3 + 2/4,   3 + 3/4
…
3 + 1/100,   3 + 2/100,   3 + 3/100,   …,   3 + 99/100.

Every rational number strictly between 3 and 4 with reduced denominator 100 or less is listed there, some repeatedly. The list has 99 × 100/2 = 4950 entries. Surround each with a subinterval 100^{-3} wide, and include the subintervals $[3, 3 + 100^{-3}]$ and $[4 - 100^{-3}, 4]$. Together, those subintervals cover less than $4952 \times 100^{-3} < 1/100$ of the length of [3, 4]. (Can you tell that if two of them are not identical, then they have to be disjoint?) The uncovered remainder of the interval is a bunch of subintervals *devoid* of rationals having reduced denominators 100 or less.

All the subintervals together constitute a partition. In that partition, the lower sum is 0, because as with Example 2 every subinterval has places where $R = 0$. In the upper sum

(LUB #1)[width #1] + (LUB #2)[width #2] + …,

some terms come from subintervals that include rationals with denominators 100 or lower. For each of those, the LUB of R is an actual maximum of 1 (as at 3/1 and 4/1) or less. Their contribution to the upper sum is therefore at most

(1)[width] + (1)[next width] +… = [sum of widths] < 1/100.

The remaining terms of the upper sum come from subintervals in which all the rationals have reduced denominators exceeding 100. In those subintervals, the maximum of R is less than 1/100, so that their contribution to the upper sum is less than

(1/100)[WIDTH] + (1/100)[NEXT WIDTH] + … < (1/100)[1].

The upper sum for R on the partition is less than 2/100.

Carry out the same argument with 10^{99} in place of 100 to see that the upper sums can be pushed arbitrarily close to 0. Since all the upper sums are positive, it follows that their greatest lower bound is 0. The upper integral matches the lower, and R is integrable.

Exercises IX.B.6b

1. Prove that Dirichlet's function is not integrable.
2. What is the value of the integral of $f(x) = x^2$ on [1, 3]?

c) looking back, looking ahead

Riemann's method is unassailable, built on the continuum properties of the reals. Still, it is worthwhile to connect it to the earlier methods of Fermat, Leibniz, Newton—even back to the method of exhaustion—all of which rely at least implicitly on the idea of limit.

The connection came from Jean-Gaston Darboux [dahr-BOO] (1842-1917). He showed that if (and only if) a function *f* is integrable, then the sums

(first value of *f*)[width of first subinterval] + … + (last value of *f*)[width of last subinterval]

have a limit as all the widths approach zero, which limit necessarily equals the common value of the Riemann upper and lower integrals.

Ahead from Riemann, the question of how continuous a function has to be to be integrable went unanswered for half a century. In his doctoral thesis of 1902, Henri Léon Lebesgue [luh-BAYG] (1875-1941) answered that what is needed is for the function to be continuous almost everywhere. [That's not just a rough description. "Almost everywhere" is the actual, precisely-defined technical phrase that applies.] Lebesgue built on the work of Borel and others to define integration in terms of **measure**, a generalization of "width." Part of Riemann's legacy is how Lebesgue's work on Riemann's integrals led to the creation of measure theory, the invariable basis for types of integrals, as a separate subdiscipline.

Oddly, Lebegue's initial interest was not what it takes for a function to be integrable, but what it takes to be differentiable. One of the things he proved is that a monotonic function—as ours tend to be over at least some intervals—is not merely continuous almost everywhere (so that it is integrable), it is *differentiable* almost everywhere.

[Lebesgue's advisor was Borel, even though the latter was just four years older. Borel had studied under Darboux. Darboux contributed in multiple fields, and eventually served as *secrétaire perpetuel* to the Academy.]

Section IX.C. New Deductive Systems

By about 1890, two brand new axiomatic systems appeared.

1. The Natural Numbers

The Weierstrass and Dedekind constructions of the real number system culminated the arithmetization of analysis. Dedekind also played an important part in the arithmetization of arithmetic. In the 1880's, he proposed a set of axioms for the system of natural numbers. Today we adopt the refinement published in the 1889 *Arithmetices principia nova methodo exposita* (*The Principles of Algebra, Presented via a New Method*) of Giuseppe Peano [peh-AH-no] (1858-1932).

a) the Peano axioms

Peano put forth the natural numbers as a set **N** of elements satisfying five axioms.

Axiom 1. Every element has a follower.

It is easier if we put the axiom in the language of functions. The axiom says that there is a **follower function** *F* that assigns to every element *n* of **N** its **follower** (or **successor**) *F*(*n*), also in **N**.

Axiom 2. Different elements have different followers.

Here the advantage of function language is evident. Axiom 2 says that *F* is one-to-one.

Axiom 3. Every element *is* a follower …

There is another half to the statement, and it is crucial. Still, it is useful to look at some examples of sets and follower functions that obey the axioms so far, yet look nothing like the natural numbers.

> Example 1. ∅, the empty set, with the "empty function," satisfies the three axioms. (Verify.)
>
> Whenever you write **universal sentences**—sentences that say "Every this …" or "All of those …"—the empty set satisfies them. It satisfies them "vacuously"; there are no elements to fail to satisfy them. For that reason, nothing in these first three axioms rules out the possibility that **N** is empty.

Example 2. $\{\pi\}$, with $F_2(\pi) = \pi$, likewise satisfies all three axioms. (Verify.)

Example 3. **R**, the set of reals, with $F_3(x) = -2x$, also fits. Again, verify that it satisfies the axioms, even though it literally turns our usual understanding of "follower" around.

The statement of Axiom 3 ends with:

Axiom 4. … with exactly one exception.

Again in the language of functions, this one says that the range of F has all but exactly one element of **N**. The function misses being *onto* by one element.

Like the identity axiom for groups (section IX.A.2a(i)), Axiom 4 requires the existence of a special element in **N**. Thereby, it rules out the empty set. That eliminates Example 1. Axiom 4 also rules out the other examples. The functions in Examples 2-3 *are* onto; neither of them allows an exception.

Let us give the exceptional natural number a name. "XCPTN" is suggestive but has too many letters. Choose instead the name "one", along with the symbol "1".

This exception has to have a follower $F(1)$. That follower cannot be the same as 1. (Explain why, remembering that the only source of explanations is the collection of axioms.) Give it a new name, say "two", denoted "2".

This second element has to have a follower $F(2)$. That follower cannot equal 1, for the same reason that $2 \neq 1$. It also cannot equal 2, because it is the follower of 2, whereas 2 is the follower of 1; 2 and 1 are different, and different elements have different followers. [Follow?] Call $F(2) = F(F(1))$ "three", denoted "3", and so on.

We have produced a list,

$1, \qquad 2 = F(1), \qquad 3 = F(2), \qquad 4 = F(3), \ldots,$

in which each entry is different from all the previous ones. *The set of natural numbers is infinite*.

Does that list all of them? Let us agree to the name **counting (up) from** k for the process of beginning with k and proceeding follower by follower:

$k, \qquad F(k), \qquad F(F(k)), \ldots.$

Counting from 1, we listed an infinity of natural numbers. In our usual picture of the natural numbers, that certainly covers them all. However, our picture does not count; only the axioms matter. Do *the axioms* guarantee that there are no other naturals? The next example shows that the answer is no.

Example 4. Take the set W of "words" you can make by stringing together one or more a's, or instead one or more a's and a terminal b, or instead an initial b followed by zero or more a's. Thus,

$a, aa, aaa, \ldots$ and $\ldots, aaab, aab, ab, b, ba, baa, baaa, \ldots$

are words. Convince yourself that those two lists comprise *all* the possible words.

On W, define a follower function by

F([word]) = [the next word in dictionary order].

Check for yourself that each of our two lists is in follower order. If you have accepted that the two lists cover all of W, you can easily see that all four axioms are satisfied.

Now, count up from the exception, the lone element a that is not a follower. You get only the all-a half of the set; you never reach the list of words with b. Axioms 1-4 are not enough.

We need one more axiom to characterize **N**.

Axiom 5. (The Induction Axiom) Suppose S is a subset of **N** with two properties:

First, $1 \in S$ (1 "belongs to" S; the symbol is going to get a lot of work);

second, if n is any element of S, then also $F(n) \in S$.

In that case, S is all of **N**.

This elegant way of formalizing mathematical induction completes the characterization of the set of natural numbers. Notice that our Example 4 violates it. In that example, the subset $S = \{a, aa, aaa, \ldots\}$ has the two properties, but does not fill up W. Still, because the example satisfies Axioms 1-4 and not Axiom 5, it is one valuable step toward proving that the axioms are **independent**: No combination (non-empty subcollection) of them implies any of the others. In different words, for each combination, you can concoct an example in which its axioms are satisfied, and not the remaining ones. In an independent set of axioms, none is redundant; that is the ideal for a deductive system. Remarkably, everything humans have come to know about arithmetic follows from this compact collection of axioms.

Theorem 1. Counting from 1 lists all the natural numbers.

> For proof, let S be the subset of **N** consisting of all the numbers in the sequence
> $1, F(1), F(F(1)), \ldots.$
> Clearly 1 is in S. Next, assume $k \in S$. Then $F(k)$ is listed next; that is how the list is made. By Axiom 5, $S =$ **N**. All the natural numbers are on the list.

The same argument establishes our principle of proof by mathematical induction (Exercise 1). From now on, we will skip the subset S in the axiom. We will seek to show, for example, that a sentence $P(n)$ about natural numbers is true for all n, rather than that the set of n for which it is true is all of **N**.

Exercises IX.C.1a

Use the five axioms to prove the statement in Exercise 1. That will establish mathematical induction as a valid method of proof, which we may thereafter use freely.

1. Let $P(n)$ be a sentence, based on the natural number n, that satisfies:
   (“base case”) $P(1)$ is true.
   (“inductive case”) Whenever $P(k)$ is true, then $P(F(k))$ follows.
   Then $P(n)$ is true of all natural n.
2. Prove that in **N**, no element is its own follower.
3. Prove that in **N**, the sequence
   $1, \quad 2 = F(1), \quad 3 = F(2), \ldots$
   has each term different from all the previous terms.

b) addition

From the axioms, we define that most elementary of operations.

Definition of addition. Fix a natural number m. By $m + 1$, we mean $F(m)$. After $m + k$ has been defined, we define $m + F(k)$ to mean $F(m + k)$.

In view of the definition, we may choose to write $m + 1$ in place of $F(m)$. Nevertheless, the function notation has advantages, including keeping the axioms in our minds. Notice, though, that the definition seems to give a separate significance to “adding to $m = 1$”, “adding to $m = 2$”, …. The separation is not objectionable; the definition attaches a meaning to every expression $m + n$. That is important: It *does* attach a meaning. It is an example of **recursive definition**. Such a description *specifies* a base case, then states not what the subsequent instances are, but how they are produced from previous cases. We encountered that kind of definition for the Fibonacci numbers (section V.B.2b): base cases $f_1 = 1 = f_2$, subsequent $f_{n+2} = f_{n+1} + f_n$. It is possible to prove from the axioms that a recursive definition yields a unique sequence of terms; we skip the proof.

The definition is a formal statement. We will soon give formal proofs to two properties of addition, then speak almost entirely *informally*. Peano’s axioms were part of a larger effort to build arithmetic from elements, as a **formal system**. A formal system encompasses a language that allows expression of definitions, axioms, and rules of inference. Such a system would be genuinely deductive, with no inference depending on assumptions that are hidden (not explicit among the axioms). We will come back to formal systems eventually, but restricting ourselves to one is laborious. We will allow ourselves the liberty to drop formality. Indeed, we have done so at least twice: Formally, the “list”

$$1, \qquad 2 = F(1), \qquad 3 = F(2), \ldots$$

calls for recursive definition (Exercise 1); and the Fibonacci recursion uses several additions, even though we are still in the process of defining addition.

Put the addition definition informally. It says that

$$m + 1 \;=\; F(m), \qquad\qquad m + 2 \;=\; F(F(m)), \qquad\qquad m + 3 \;=\; F(F(F(m)), \ldots.$$

In words, the way you get $m + n$ is to count n numbers up from m. Did you ever see a child—or even an adult—add by counting on his fingers? If yes, did it make you think that he could not add? Observe now that adding is *precisely* what he was doing. If you see “5 + 3” and immediately decide “8”—even picturing a figure-8 in your mind—then you are not adding. You are not processing an algorithm. You are *recalling* or *accessing* a memorized fact. The counting child simply has not stored that datum.

On the formal track, let us establish addition’s most basic properties.

Theorem 1. Addition is associative: Suppose k, m, and n are in **N**. Then

$$(k + m) + n \;=\; k + (m + n).$$

The proof method of choice is obvious; induction is the only weapon we have.

$n = 1$ case: $(k + m) + 1 \;=\; F(k + m)$ (by the definition of adding 1)
$\;=\; k + F(m)$ (inductive part of addition definition)
$\;=\; k + (m + 1)$ (definition of adding 1).

That establishes the $n = 1$ case.

inductive case: Assume $(k + m) + n \;=\; k + (m + n)$. Then
$[k + m] + F(n) \;=\; F([k + m] + n)$ (definition of addition)
$\;=\; F(k + [m + n])$ (by assumption)
$\;=\; k + F([m + n])$ (addition)
$\;=\; k + [m + F(n)]$ (same).

That establishes the inductive case, completing the proof by induction of associativity.

In the middle of the argument, we concluded from the assumption

$$[k + m] + n \;=\; k + [m + n]$$

that $\quad F([k + m] + n) \;=\; F(k + [m + n])$.

The reason is nowadays built into the *definition* of function: $x = y$ forces $F(x) = F(y)$. Peano explicitly listed the principle among the totality of axioms, as part of the project to assemble a formal system.

Theorem 2. Addition is commutative: If m and n are natural, then $m + n \;=\; n + m$.

case $n = 1$: We will prove that $m + 1 \;=\; 1 + m$ by “double induction,” induction on m within the induction on n.

case $m = 1$: That one says $1 + 1 \;=\; 1 + 1$; sounds reasonable.

inductive case for m: Assume $m + 1 \;=\; 1 + m$. Then
$F(m) + 1 \;=\; F(F(m))$ (definition of adding 1)
$\;=\; F(m +1)$ (same)
$\;=\; F(1 + m)$ (assumption)
$\;=\; 1 + F(m)$ (inductive part of addition definition).

That establishes the inductive case; it proves by induction on m that $m + n \;=\; n + m$ in the case we are considering, $n = 1$.

inductive case for n: Assume that $m + n = n + m$. Then

$$\begin{aligned} m + F(n) &= m + (n + 1) && \text{(adding 1)} \\ &= (m + n) + 1 && \text{(associativity)} \\ &= 1 + (m + n) && \text{(the } n = 1 \text{ case)} \\ &= 1 + (n + m) && \text{(assumption)} \\ &= (1 + n) + m && \text{(associativity)} \\ &= (n + 1) + m && \text{(the } n = 1 \text{ case)} \\ &= F(n) + m && \text{(adding 1).} \end{aligned}$$

That establishes the n-induction, proving for all natural m and n that $m + n = n + m$.

Exercises IX.C.1b

1. Give a formal definition of our “list”

   $1, \quad 2 = F(1), \quad 3 = F(2), \quad 4 = F(3), \ldots.$

2. Prove that addition allows cancellation:

   If $k + n = m + n$, then necessarily $k = m$.

3. a) Write a formal definition of $F^n(m)$, what we write informally as $F(F\ldots(F(m)))$.

   b) Use your definition to prove that

   $m + n = F^n(m)$.

   We spoke that equality as, “You get $m + n$ [by counting] n numbers up from m.”

c) order and arithmetic

To define an order (relation), remember that the list

$1, \quad 2 = F(1), \quad 3 = F(2), \ldots$

comprises all natural numbers. There, each term differs from all the previous ones. In other words, every element of **N** appears exactly once on the list. Accordingly, if m and n are distinct natural numbers, then one of them appears before the other. We will say that m **is smaller than** n, and write $m < n$, if m appears before n on the list. (Make a formal definition in Exercise 1.) Equivalently, we say n **is greater than** m and write $n > m$. We add $j \leq m$ to signify that either $j = m$ or $j < m$; similarly for $k \geq n$.

We can now state some familiar properties of the order. Formal proofs are possible, but mostly we either skip them completely or leave formal (or informal) argument to the exercises.

Theorem 1. For natural numbers m and n, the following are equivalent:

a) $m < n$.

b) You can count from m up to n.

c) You *cannot* count from n up to m.

d) There exists a natural number k such that $n = m + k$.

In view of (d), we can now define **subtraction**: If $m < n$, we write $n - m$ to signify the one natural y such that $n = m + y$. (See Exercise 3.) Subtraction has familiar properties.

Theorem 2. For natural k, m, and n:

a) If $k > m + n$, then

$[k - m] \qquad \text{and} \qquad [(k - m) - n]$

are both defined, and

$$k - (m + n) = (k - m) - n.$$

b) If $k > m$ and $m > n$, then

$k - (m - n)$

is defined, and

$$k - (m - n) = (k - m) + n.$$

For part (a), assume $k > m + n$. By Theorem 1(d), $m + n > m$. By transitivity (adapt Exercise 2), $k > m$. Therefore $k - m$ is defined. Further, $k - m > (m + n) - m$ (Exercise 5a). That says $k - m > n$, which implies that $[(k - m) - n]$ is defined.

Write $k - (m + n) \ = \ x$. Then

$$\begin{aligned} (m + n) + x \ &= \ k && \text{(definition of subtraction)} \\ &= \ m + (k - m) && \text{(same)} \\ &= \ m + (n + [(k - m) - n]) && \text{(same)} \\ &= \ (m + n) + [(k - m) - n] && \text{(associativity).} \end{aligned}$$

By cancellation (Exercise IX.C.1b:2),

$$k - (m + n) \ = \ x \ = \ (k - m) - n.$$

For part (b), do Exercise 5b-c.

The other operations get the definitions you would expect. Multiplication is defined by

$mn = m + m + \ldots + m$ (n summands; compare Exercise 6).

Division is its partial inverse. If there is a k such that $m = kn$ (for which we use the usual language, like “n divides m”), then there will be just one such k (Exercise 6c), and $m \div n$ means k.

Exercises IX.C.1c

This is a big set of exercises that are either long or sophisticated. Pick your spots. Except where formality is specifically required, informal proof will suffice. For any proof, you may refer to any previous ones.

1. a) Give a formal definition of $m < n$. (Hint: Work as in Exercise IX.C.1b:3a.)
   b) Use the definition to prove that if $m < n$, then there exists a natural k such that $n = m + k$.
2. Show that $\leq$ is a **partial order**; that is, it is:
   a) **reflexive**: $m \leq m$ for every m.
   b) **antisymmetric**: If $m \leq n$ and $n \leq m$, then $m = n$.
   c) **transitive**: If $k \leq m$ and $m \leq n$, then $k \leq n$.
3. Prove that if there is some y with $n = m + y$, then there is only one such y.
4. Prove that subtraction is not associative.
5. a) Show that for natural x, y, and z,
   if $x > y > z$, then $x - z > y - z$.
   (Why does “$x > y > z$” makes sense?)
   b) Show that if $m - n$ is defined, then $m - n < m$.
   c) Prove part (b) of Theorem 2.
6. a) Give a formal definition of multiplication.
   b) Prove that mn is always greater than m, except that $m1 = m$.
   c) Prove that multiplication allows cancellation: If $kn = mn$, then $k = m$.
7. Prove that multiplication is:
   a) associative: $(km)n \ = \ k(mn)$
   b) commutative: $mn = nm$
   c) distributive over addition: $k(m + n) \ = \ km + kn$
   d) “compatible” with >: If $m > n$, then $km > kn$.
   e) distributive over subtraction: If $m > n$, then $km - kn$ is defined, and $k(m - n) \ = \ km - kn$.

8. Show that in **N**, 3÷2 is not defined, but 4÷2 is.
9. A partial order on a set (refer to Exercise 2) is called a **total order** if every pair of elements is **comparable**: Given *x* and *y* (not necessarily distinct), either *x* is related to *y*, or *y* is related to *x*.
   a) Show that on **N**, ≤ is a total order.
   b) Show that on **N**, "*m* divides *n*" defines a (relation that is a) partial order.
   c) Show that "*m* divides *n*" does not define a *total* order.

d) the well-ordering and other principles

We end our discussion of the axiomatization of **N** with three principles. It is easy to prove them from the five axioms. It is harder to prove formally that in the presence of Axioms 1-4, they all imply the induction axiom. In other words, they are actually equivalent to Axiom 5. We will settle for an informal chain of reasoning leading to the equivalence. We list first the most important, the one we have invoked most frequently. The other two are worth expressing because they are (fairly) well-known and useful.

The Well-Ordering Principle. Under the five axioms, every nonempty subset of **N** has a least element.

Suppose T is a subset of **N**. (Peano introduced the notation $T \subseteq \mathbf{N}$.) We need to show that if T is nonempty, then T has a least element. We prove instead the contrapositive, by the induction axiom.

> Suppose $T \subseteq \mathbf{N}$ has no least element. Look at the set S of natural numbers n such that every number smaller than or equal to n is *outside* T. In symbols,
> $S = \{n\text{: } 1, 2, \ldots, n \text{ are not in } T\}$.
>
> First, $1 \in S$. The only natural number smaller than or equal to 1 *is* 1, and it has to be outside T. If it were in T, then it would be the smallest element there; T does not have a smallest. Second, if $k \in S$, then $F(k)$ must likewise be in S. If $k \in S$, then $1, 2, \ldots, k$ are all outside T. Accordingly, $F(k)$ must also be outside T: If $F(k)$ were in T, *it* would be the least element there. That puts $F(k)$ in S.
>
> We see that S satisfies the hypotheses of the induction axiom. Therefore S is all of **N**. That is, all natural numbers are outside T. We have proved T is empty.

If we rely on just Axioms 1-4, then we lose the recursive definition of the list
$1, \quad 2 = F(1), \quad 3 = F(2), \ldots$
(Exercise IX.C.1b:1) and the proof that the list covers of all **N** (Theorem 1 in section IX.C.1a). Still, we can informally define what it means for m on that list to be smaller than n. One of them has to come first. If the first is m, then $n = F(F\ldots(F(m)))$, and we use that to define $m < n$. With that in mind, we can prove that the well-ordering principle implies the next principle (Exercise 1a).

The Principle of Complete Induction. Suppose $P(n)$ is a sentence, about the natural number n, with these two properties:
(base) $P(1)$ is true.
(inductive) If $P(1), P(2), \ldots, P(k)$ are all true, then $P(k + 1)$ is true.
Then $P(n)$ is true for every $n \in \mathbf{N}$.

This principle describes another approach to proof by induction. You might think that it is harder to apply than our usual method. Actually the opposite is true. This version does not *require* more information. Instead, it *allows you to assume* more information. For that reason, it is sometimes easier to apply, as in proving the next statement (Exercise 1b).

The Principle of Infinite Descent. Suppose $Q(n)$ is a sentence, about the natural number n, such that:
If $Q(n)$ is true, then there is a *smaller* natural m for which $Q(m)$ is likewise true.
Then $Q(n)$ is never true, for any number n.

We may put the principle informally by saying that **N** does not allow infinite descent. Our first encounter with it happened in Egypt; it underlies the explanation why the "biggest-fit method" ends with a fraction's decomposition into unit fractions (Exercise II.A.4:6b). For a less ancient example, from a practitioner who frequently invoked the principle, view Fermat's proof (section VII.A.4f(iv)) that

$$x^2 + y^4 = z^4$$

has no integer solution.

> To paraphrase Fermat, let $R(n)$ represent the sentence:
>
> For the natural number n, there exist integers x and y with $x^2 + y^4 = n^4$.
>
> Fermat established that this sentence satisfies the hypothesis of the principle: If the equation has a solution for one natural number n, then it also has a solution for some smaller m. Therefore by the principle, $R(n)$ is always false.

Look back at Example 4 in section IX.C.1a. There, we found that the set

$W = \{a, aa, aaa, \ldots$ together with $\ldots, aab, ab, b, ba, baa, \ldots\}$

satisfies Axioms 1-4. We saw further (via Theorem 1 there) that it violates Axiom 5. Observe now that it also violates the principle of infinite descent; it allows infinite descent along the *b*-line.

We saw that the induction axiom implies the well-ordering principle. Well-ordering implies complete induction, which implies the infinite descent principle. All of those assume Axioms 1-4, and the last two are Exercise 1. Below, we will prove that the principle of infinite descent implies the induction axiom. That chain of inferences,

Axiom 5 implies well-ordering	(proved above)
implies complete induction	(Exercise 1a)
implies infinite descent	(1b)
implies Axiom 5	(below),

establishes that the four principles are equivalent.

> Assume the principle of infinite descent. To prove the induction axiom, assume further that S is a subset of **N** with the properties:
>
> (base) $1 \in S$;
>
> (inductive) Whenever $k \in S$, then also $F(k) \in S$.
>
> Look at the sentence $R(n)$ that says:
>
> $n \notin S$ (n is not an element of S).
>
> Suppose $R(n)$ is true, meaning n is not in S. First, n cannot be 1. $R(1)$ is not true:
>
> What $R(1)$ says is $1 \notin S$; that is false by the base condition.
>
> Second, since $n \neq 1$, n is a follower, say $n = F(m)$. The statement $R(m)$ has to be true:
>
> What $R(m)$ says is $m \notin S$. If that were false, meaning $m \in S$, then the inductive condition would force $n = F(m) \in S$, contrary to the assumption that $R(n)$ is true.
>
> From the assumption that $R(n)$ is true, we have concluded that $R(m)$ is true, m being smaller than n. By the overlying assumption, namely the principle of infinite descent, the sentence $R(n)$ is false for all n. That is, every natural number is in S; $S = $ **N**. That proves the induction axiom.

Exercises IX.C.1d

1. Assuming the first four Peano axioms, prove that:
   a) The well-ordering principle implies the principle of complete induction.
   b) The principle of complete induction implies the principle of infinite descent.

2. Peano held a university post at Turin (and later at the Military Academy, as had Lagrange). Why were mathematicians in the nineteenth century more likely than in previous centuries to be academics?

e) evolution of "axioms"

For two thousand years after Euclid, the word "axiom" held one meaning. An **axiom** was a statement whose truth was so plain as to be undeniable, something nobody could doubt. A good example is, "When equals are added to equals, the results are equals." We have seen from the creations of Bolzano, Cauchy, Galois, Weierstrass, Dedekind, and Peano that before the nineteenth century ended, the meaning had become "working assumption." A working assumption is a statement to which the reader is asked to agree for the time being, presumably because some mathematical profit lies in acceptance. The latter meaning was what Euclid had given to the word "postulate." You would need a great sense of humor to think of "Euclid's postulate" as plainly true. Euclid's geometry showed what you could conclude if you chose to accept the postulate. As we will see later, the creators of "non-Euclidean geometry" showed what you could conclude if you chose to deny it.

Remarkably, the evolution in the meaning of "axiom"—from undeniable truth to assumption worth considering—was reflected in two of the most important writings in the life of the United States. In 1776, Thomas Jefferson's Declaration included the words, "*We hold these truths to be self-evident,* that all men are created equal …". Eighty-seven years later, Abraham Lincoln's "few appropriate remarks" at the consecration of the Gettysburg cemetery included, "Now we are engaged in a great civil war, *testing whether that nation, or any nation so conceived* and so dedicated, *can long endure*."

[Italics added. I heard the comparison from the late Alvin Hausner. Read about how the Gettysburg Address is grounded in the Declaration of Independence and in Pericles's Funeral Oration in *Lincoln at Gettysburg,* by Garry Wills.]

2. Sets

The second new axiomatic system covered the theory of sets. It grew from the work of Georg (Ferdinand Ludwig Phillip) Cantor (1845-1918) in the last quarter of the nineteenth century.

Cantor's first interest was number theory. Then at the behest of Heine, he successfully tackled the question of uniqueness of trigonometric series. Recall (section VIII.B.4a) that if a function is given by a power series, then that series has to be the Taylor series. Cantor showed that the Fourier series is the only *trigonometric* series that can give a function. That result had eluded all comers, including Dirichlet and Riemann. While attacking the problem, Cantor became interested in infinite sets.

Cantor tried to axiomatize set theory. Much as mathematicians had been willing to work with intuitive ideas of "function" and "continuity," so had they been satisfied with the intuitive notion of "set." Consider:

A set is a collection of objects with some defining characteristic.

You can see that the last sentence is hopeless as a definition, simply substituting "collection" for the target word plus adding such undefined words as "characteristic."

Cantor's system proved to be too broad. He himself saw the logical difficulties that arise if you allow sets that are "too big," like the set *S* of all sets. That set would have *S* itself as a member, inside of which member there would be a member *S*, inside of which …. Among other contradictions, an elementary and serious one is the **Russell paradox**.

> There are sets that are elements of themselves. If we allow *S* above, we have an example. The set of infinite sets is another, since it has an infinity of members. On the other hand, the set **N** of natural numbers is not a natural number, is therefore not one of its own members. Looking now at the set *U* of sets that are not members of themselves, we see that
>
> *U* is a member of *U* iff *U* is not a member of *U*.
>
> (See also Scientific American.)

It took decades for the work of Bertrand Russell, Ernst Zermelo, and Abraham Fraenkel to restructure the Cantor axioms to avoid such contradictions.

The part of Cantor's theory we will study has to do with numbers of elements in sets. This part astounded the mathematical world because it established that there is not just one infinity, that some infinite sets are more infinite than others. Irrespective of the paradoxes, this part of his work was subjected to tremendous (and vicious) criticism on mathematical and philosophical grounds. (Read **Boyer** on Kronecker's reaction.) It was years before the mathematical world came to agree on the validity—indeed, brilliance—of Cantor's work.

a) numerousness

In developing the idea we called "number of elements," Cantor found it easier to begin with "equal numbers of elements." Einstein did something similar later when he found that the idea of *time* was easier to describe in terms of "equal-time-ness," or *simultaneity*.

(i) sets and functions

We must recall some vocabulary from sets and functions.

We have already much used the intuitive ("naïve") concepts of "set" and "belonging to" (same as "being a member or element of") a set, and will continue to use them. A set *T* is a **subset** of set *S* if every member of *T* belongs also to *S*. If *S* has all the members of *T plus others*, then we say *T* is a **proper subset** of *S*.

Given sets *U* and *V*, a **function from** [or **mapping**] *U* **to** (or **into**) *V* assigns to *every* element of *U* a single element of *V* (Dirichlet's definition). If the function *f* assigns to *u* in *U* the element *v* in *V*, then we write $v = f(u)$ (per Lagrange) and say *v* is the **image** of *u* under *f*. If different members of *U* necessarily have different images in *V*—if

$u \neq t$ forces $f(u) \neq f(t)$—

then *f* is **one-to-one**. If every element *v* in *V* is the image $v = f(u)$ of some *u* in *U*, then *f* is **onto**. If *f* is both one-to-one and onto, then *f* is called a **one-to-one correspondence**. (View Exercise 1.)

(ii) equinumerousness

Now we can give Cantor's definition, from around 1874. The sets *U* and *V* **are equivalent,** or **have the same cardinality**, if there exists a one-to-one correspondence between *U* and *V*. (Do Exercise 2 to see that the relation of equivalence really is symmetric. Cantor also used the German word for "power" in place of "cardinality." When *U* and *V* are equivalent, we will informally say that they *are equally numerous* or *have the same number of elements*. Because we have to say "one-to-one correspondence" so often, we will abbreviate it to "correspondence.")

Consider the set {a, e, i, o, u} of English lower-case vowels, the set of complex roots of $z^5 - 1$, and the set of Peano axioms. Those are collections of very different kinds of objects, but they are all equivalent to the subset {1, 2, 3, 4, 5} of **N**. We could define the **cardinal number** 5 or "fiveness" as the quality that all such equivalent sets share.

(iii) infinite numerousness

Look at the correspondence from **N** to the set *E* of *even* natural numbers described by

$1 \leftrightarrow 2, \qquad 2 \leftrightarrow 4, \qquad 3 \leftrightarrow 6, \qquad \ldots.$

Already by 1600, Galileo had observed that it suggests that *E* is as numerous as the complete **N**. Clearly the latter is a bigger set in the sense that it has all the elements of the former and *more*; in other words, *E* is a proper subset of **N**. But matching them up as in the arrow diagram, we cannot escape the idea—which Cantor's definition now makes precise—that *E* has as many elements as **N**.

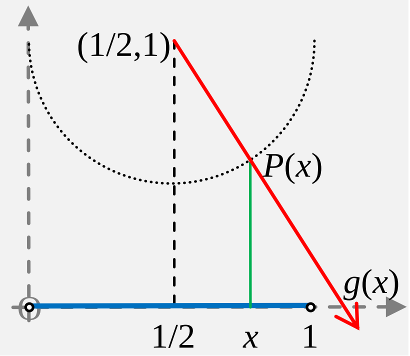
Galileo seemed to view it as a curiosity. Bolzano saw a paradox. Dedekind, instead, realized that it captures the very nature of the infinite. Accordingly, Dedekind gave this definition (1872): A set is **infinite** if it can be put in correspondence (Cantor's later language) with a proper subset of itself.

Consider **R**. Since it contains a copy of **N**, we certainly think of it as infinite. Let us put it to Dedekind's definition.

In the figure at right, we lay out the real numbers along the *x*-axis and highlight (blue) the **open interval** (0, 1) of real *x* with

$0 < x < 1$.

[European notation for the interval without its endpoints is "]0, 1[". That has great advantage in avoiding conflict with *coordinate* notation like the figure's (1/2, 1).] Draw the lower half (shown dotted) of the circle of radius 1/2 centered at (1/2, 1). Given *x* strictly between 0 and 1, draw the vertical (green) from (*x*, 0) on the axis to the point *P*(*x*) on the circle. Then draw the half-line (red) from (1/2, 1) through *P*(*x*) and onward through the axis, meeting the axis at *g*(*x*). That red ray sweeps out the entire *x*-axis without repetition; *g* is a correspondence. Therefore **R** satisfies Dirichlet's definition.



Exercises IX.C.2a

1. Show that as functions from **R** to **R**:
   a) $f(x) = x^2$ is neither one-to-one nor onto.
   b) $g(x) = e^x$ is one-to-one but not onto.
   b) $h(x) = x^3 - 3x$ is onto but not one-to-one.
   b) $H(x) = x^3$ is both.
2. a) Show that equivalence of sets is a symmetric relation: If there is a correspondence from *U* onto *V*, then there is a correspondence from *V* onto *U*.
   b) Show that equivalence is an equivalence relation (defined in Exercise VIII.C.2a:3).
3. Write a formula that gives a correspondence between the closed intervals [3, 7] and [1, 11].
4. In the correspondence $x \to P(x) \to g(x)$ given by the last figure:
   a) Find the formula for the *y*-coordinate of *P*(*x*).
   b) Find a formula for *g*(*x*).
5. In view of the correspondence between (0, 1) and **R**, find a correspondence between the closed interval [0, 1] and **R**. (A verbal description will suffice. Hint: $H(n) = n + 2$ describes a correspondence between **N** and {3, 4, 5, …}.)
6. Show that if *T* is a subset of *S* and *T* is infinite, then *S* is infinite.

b) how many

(i) unique number

To define "number of elements," the obvious starting point is to define "oneness," the property of having one element. We will say the set S **has exactly 1 element** to mean that S is nonempty and that, if a and b belong to S, then $a = b$. (The use of "1," representing the lone non-follower in **N**, is deliberate.) If that defining sentence looks strange, remember that things can have different names. In a field,

$u(v + w)$ and $uv + uw$

name equal elements.

The definition puts oneness in terms of the more elementary notion of equality, which perhaps has to be left undefined. Equality can be used in the definitions of *function* (if $a = b$, then $f(a) = f(b)$), *one-to-one function* (if $f(a) = f(b)$, then $a = b$), and 1 (if a and b are non-followers in **N**, then $a = b$). [The idea is to avoid a circular definition of oneness. If you can convince me that this approach is circular anyway, or instead that it is not, I will be happy to put your argument in this place.]

To define "number of elements" recursively, assume we know what it means for a set to have the natural number n of members. Then the set S **has *n* + 1 members** if it has a **singleton** (1-member) subset T whose complement T^* (comprising the members of S outside T) has n members.

Theorem 1. S has n members iff it is equivalent to $\{1, 2, \ldots, n\}$.

It will be handy to abbreviate $\{1, 2, \ldots, n\}$ by S_n. We call S_n a **segment** of **N**.

As long as we are counting by means of such natural number sets, the natural method of proof is induction. We will argue that if S has n members as just defined, then there is a correspondence between S and S_n. For the converse of the theorem, we work the base case, and leave evidence for the inductive case to Exercise 1.

(base case) Suppose S has 1 member. Set $f(a) = 1$ for every a in S. That defines a function from S to $\{1\} = S_1$, because if $a = b$ in S, then

$$f(a) = 1 = f(b).$$

The function is one-to-one, because if $f(c) = f(d)$, then $c = d$; the latter equality is automatic in S. The function is onto, because S has *some* member e, and that member has $f(e) = 1$. Therefore f is a correspondence between S and S_1.

(inductive case) Assume that every set of 100 elements is equivalent to S_{100}. Suppose S has 101 elements. By definition, S contains a singleton T whose complement T^* has 100 members. By the base case, there is a correspondence h_1 between T and $\{1\}$. By the inductive hypothesis, there is a second correspondence h_2 between T^* and S_{100}. Define h on S by

$$\begin{aligned} h(s) &= h_1(s) && \text{if } s \in T \\ &= h_2(s) + 1 && \text{if } s \notin T. \end{aligned}$$

Then h maps S one-to-one onto S_{101}: The lone image 1 from T does not match any of the images 2-101 from T^*, and 2-101 come from different members of T^*. Hence S is equivalent to S_{101}.

Conversely, suppose g is a correspondence from $\{1\}$ to S. First, S is nonempty, because $A = g(1)$ is a member. Next, if B is also a member of S, then B is an image $g(b)$, because g is onto. This b is another non-follower in $\{1\}$, so $b = 1$. Because g is a function, $B = g(b)$ must equal $A = g(1)$. Therefore S has 1 member.

Now we see the natural numbers not just as creatures of Peano, but as measures of "how many." We must next show that a set *determines* its number of elements, that S cannot have m and $n \neq m$ members.

Theorem 2. A function from S_{n+1} to S_n cannot be one-to-one.

Notice that Theorem 2 implies that you cannot map S_{n+k} one-to-one to S_n, no matter what natural number k is. [Proof doesn't even need induction] The last guarantees that S cannot have both m and $n \neq m$ members. If that were so, you could map $S \leftrightarrow S_m$ and $S \leftrightarrow S_n$, both one-to-one onto. Then from symmetry and transitivity (Exercise 2b above), we would have a correspondence $S_m \leftrightarrow S_n$.

Theorem 2 amounts to a formal statement of the **pigeonhole principle**. See the informal statement at the end of section VII.A.4f(ii). As expected, our evidence is inductive.

(base case) Let f map $\{1, 2\}$ into $\{1\}$. The latter set has just one member, so $f(1) = f(2)$. We know 1 and 2 are unequal elements of $\{1, 2\}$, because 2 is a follower. Hence f is not one-to-one.

(inductive case) We assume that no function from S_{100} to S_{99} is one-to-one, and let g map S_{101} to S_{100}. The argument below is abstruse, but it amounts to something simple: g cannot be one-to one, because either $g(101)$ matches one of the images $g(1), \ldots, g(100)$, or one of those matches another.

If $g(101)$ matches any of the images $g(1), \ldots, g(100)$, then g is not one-to-one.

Suppose instead that $g(101)$ does not match another image, say $g(101) = 52$ and 52 is not any of $g(1), \ldots, g(100)$. In that case, those 100 images are red numbers from 1 to 51 or 53 to 100. Define function H from $S_{100} = \{1, \ldots, 100\}$ to S_{99} $\{\underline{\underline{1}}, \ldots, \underline{\underline{99}}\}$ by the formula

$$\begin{array}{llll} H(i) & = \underline{g(i)} & \text{if} & 1 \leq g(i) \leq 51 \\ & = \underline{g(i)} - 1 & \text{if} & 53 \leq g(i) \leq 100. \end{array}$$

Check for yourself that H is a function and maps 1-100 to $\underline{\underline{1}}$-$\underline{\underline{99}}$.

By the inductive hypothesis, H is not one-to-one. That means there are j and $k \neq j$ in S_{100} such that

$$H(j) = H(k).$$

That equality forces $g(j) = g(k)$. That tells us g is not one-to-one, completing the induction.

Another consequence of Theorem 2 is that you cannot map S_m one-to-one into a proper subset of itself. That means S_m fails Dedekind's definition, as does every set equivalent to it. We therefore describe them as **finite** ("uninfinite"?). (We call the empty set finite as well. Notice that we have confirmed the statement from section IX.A.1d(i) that a permutation, defined as a one-to-one mapping of S_m into itself, is necessarily onto.)

(ii) counting elements

In any nonempty set, we can **count** (map one-to-one onto some segment of **N**) *some* elements. Given set S, we may choose some element; call it a_1 to indicate the correspondence $\{a_1\} \leftrightarrow \{1\}$. If there are other elements, we choose $a_2 \neq a_1$, and correspond $\{a_1, a_2\} \leftrightarrow \{1, 2\}$. (That such "choices" are possible is so important that it constitutes an axiom in the later development of set theory. We have no hope of delving into it here.) It is possible that this process eventually runs out of candidates, with $a_1, a_2, \ldots, a_n$ distinct and exhausting S. In that case S is finite and has n elements.

Alternatively, the process might never terminate. In that case, $\{a_1, a_2, \ldots\}$ is a subset of S equivalent to **N**. We may write:

Theorem 3. A set is infinite iff there is a correspondence between some subset of it and **N**.

Remember that "infinite" means satisfying Dedekind's definition.

Assume S is infinite. Then the choosing process above can never end. If it did, then we would find

$$S = \{a_1, a_2, \ldots, a_n\},$$

equivalent to S_n. Since S maps one-to-one onto a proper subset, there would be a corresponding map of S_n onto one of *its* subsets. We ruled that out based on Theorem 2. Therefore the choosing goes on without end, and produces a subset of S equivalent to **N**.

Conversely, assume there is such a subset $T = \{b_1, b_2, \ldots\}$. Define f on S by

$$f(s) = \begin{cases} s & \text{if } s \text{ is outside } T \\ b_{n+1} & \text{if } s \text{ is the element } b_n \text{ inside } T. \end{cases}$$

Observe that $f(s)$ is uniquely defined for every s; f is a function on S. It is one-to-one, because images from T are distinct; images from the complement T^* are distinct; and an image from T cannot match an image from T^*, since they are in T and T^* respectively. No member of S has b_1 as image. Consequently, f is a one-to-one mapping of S onto a proper subset of S. That makes S infinite.

Exercises IX.C.2b

1. Assuming that every set equivalent to S_{100} has 100 members, show that every set equivalent to S_{101} has 101 members.
2. Show that if S has $m + n$ elements, then there is a subset T of S having m elements and complement T^* with n elements.

c) countable sets

We saw that you can count some elements in any nonempty set. If you can count them all—if there is a correspondence between S and either some segment or all of **N**—then we say S is **countable** (or **denumerable**). It is usual to, well, count the empty set as countable.

It is perhaps not surprising that the set **Z** of integers, despite having what looks like twice the population of **N**, is countable. The function given by

$$g(i) = \begin{cases} 2i + 2 & \text{if } i \geq 0 \\ -2i - 1 & \text{if } i < 0 \end{cases}$$

(nonnegatives go to the evens, negatives to the odds) is a correspondence between **Z** and **N**.

Such correspondences are typically not easy to establish. It is better to rely on the next result.

Theorem 1. If S is an infinite set and either:

a) There is a function mapping **N** (not necessarily one-to-one) *onto* S, or

b) There is a function mapping S one-to-one into (not necessarily onto) **N**,

then S is countable (or for more specificity, **countably infinite**).

a) Assume h maps **N** onto S. Write $a_1 = h(1)$. We know S has other members, so $h(1)$ is not the only image: There are numbers $k \neq 1$ for which $h(k) \neq a_1$. By the well-ordering principle, there is a smallest such number $k_2 \geq 2$. That means

$$a_1 = h(1) = h(2) = \ldots = h(k_2 - 1) \quad \text{but} \quad a_2 = h(k_2) \neq a_1.$$

(Here, a_1 may well be the image of numbers beyond k_2.) Again, S has to have members other than a_1 and a_2. There must be numbers whose images are not those two, and so there is a smallest such number $k_3 \geq 3$. Accordingly,

each of $h(1), h(2), \ldots, h(k_2 - 1)$ is a_1,
each of $h(k_2), h(k_2 + 1), \ldots, h(k_3 - 1)$ is either a_1 or a_2,
but $a_3 = h(k_3)$ is different from both a_1 and a_2.

The process goes on indefinitely. We see that h maps every natural from 1 to $k_n \geq n$ to $a_1, \ldots, a_n$. Therefore $a_1, a_2, \ldots$ are all the images under h. Since h is onto, that implies $\{a_1, a_2, \ldots\} = S$. The set is countable.

b) Exercise 1.

Think of Theorem 1a as a statement that if you can make a natural-numbered *list* of *all* the members of a set, then the set is countable. (Clearly the converse is also true.) With that viewpoint, we demonstrate that the infinite infinity of nonnegative rational numbers is nevertheless countable.

> Look at the following list of rationals, arranged in rows by sum of numerator and denominator:
>
> | | |
> |---|---|
> | row #1 | 0/1 |
> | row #2 | 0/2, 1/1 |
> | row #3 | 0/3, 1/2, 2/1 |
>
> ….
>
> Every nonnegative rational is there. For example, 705/453 is in row #(705 + 453), at position 706 from left. (Indeed, every rational is there multiple times; do Exercise 2.) The first 1157 rows have 1157(1158)/2 entries, so 705/453 is the [1157(1158)/2 + 706]'th rational on the list. The array defines a function from **N** *onto* the set of nonnegative rationals. By Theorem 1a, those rationals constitute a countable set. (Do Exercise 3.)

Next example is a greater surprise. Consider the set Σ of finite sequences of natural numbers. It includes the one-term sequences (1), (2), …; and the squared infinity of ordered pairs (m, n); and the cubed infinity of triples …. Still, Σ is countable.

> Write a list of prime numbers $p_1, p_2, \ldots$. It does not matter if they are out of size-order or if some primes are missing, provided they are distinct. (Why is such an infinite list possible?) Given the sequence $s = (m_1, m_2, \ldots, m_k)$, define
>
> $$h(s) = p_1^{m_1} p_2^{m_2} \ldots p_k^{m_k}.$$
>
> This h assigns to s a unique natural number. If $s \neq t$, then $h(s)$ and $h(t)$ are different prime-power factorizations, are therefore unequal natural numbers. Thus, h maps Σ one-to-one into **N**. By Theorem 1b, Σ is countable. (Compare Exercise 4.)

Notice that the same argument, restricted to just the ordered pairs (m, n), would have demonstrated that the rationals are countable.

It is customary to symbolize the number of natural numbers—the cardinality of **N**—by $\aleph_0$. The symbol is read "aleph null," using the Hebrew letter *aleph*. From the last example, we see that $\aleph_0$ is big enough to make, not just

$$\aleph_0 = \aleph_0^2 = \aleph_0^3 = \ldots,$$

but in fact

$$\aleph_0 = \aleph_0 + \aleph_0^2 + \aleph_0^3 + \ldots.$$

[Maybe we should say it "is small enough."]

Our last example is the set of algebraic real (or even complex) numbers. Recall (section IX.A.2c(iv)) that a real or complex number is called **algebraic** if it is a root of some polynomial with rational coefficients. Every such polynomial $p(x)$ has a unique **standard form**

$$p(x) = r_n x^n + r_{n-1} x^{n-1} + \ldots + r_1 x + r_0,$$

with the exponents decreasing and any missing powers written in with zero coefficients. The function given by

$$f(p) = (r_n, r_{n-1}, \ldots, r_1, r_0)$$

maps the set P of those polynomials one-to-one onto the set Q of finite sequences of rational numbers. Because the rationals are countable (Exercise 3), there is a correspondence between Q and Σ (finite sequences of *naturals*). We just saw that there is a correspondence between Σ and **N**. The composition of all those gives a one-to-one function from P to **N**. By Theorem 1b, the set of polynomials is countable. By Exercise 5, the set of algebraic numbers is countable.

Exercises IX.C.2c

1. Prove that if function H maps the infinite set S one-to-one into **N**, then S is countable.
2. We stated that 705/453 is at position 706 in row #1158 in our array. Where else is it?

3. Show that there is a correspondence between **N** and *all* of the rationals.
4. Show that the class (set) of finite subsets of **N** is countable.
5. Show that the algebraic numbers form a countable set. (Partial hint: For every algebraic number, there is a unique *monic minimal* polynomial [section IX.A.2c(v)].)

---

d) uncountable sets

Just as "acoustic guitar" exists only because there are now other kinds, so does inventing the adjective "countable" make sense only if some sets are not. The followers of Kronecker and Wittgenstein would have dismissed the idea of "actually infinite" sets, as opposed to infinite in some unreachable limiting sense; they would have considered it, at best, nonsense. When Cantor proposed that there exist not merely infinite sets, but different levels of infinity, they just about called the idea unholy. The opposition harmed Cantor's academic career. However from around 1900, and with the support of Hilbert, the mathematical validity of Cantor's work has been beyond doubt.

(i) the reals

The fundamental uncountable set is the set of real numbers. It has as many members as the open interval (0, 1). To see the latter's cardinality, represent each of its reals as a decimal.

> Take the numbers 1/8 and √2/2. For each, look among 0/10, 1/10, …, 9/10 for the last one smaller than it. Thus, we find numerators $a_1 = 1$ and $b_1 = 7$ with
>
> $$a_1/10 \;<\; 1/8 \;\le\; 2/10 \qquad \text{and} \qquad b_1/10 \;<\; \sqrt{2}/2 \;\le\; 8/10.$$
>
> Next, find the last of
>
> $$a_1/10 + 0/100, \qquad \ldots, \qquad a_1/10 + 9/100$$
>
> smaller than 1/8, and similarly with $\sqrt{2}\sqrt{2}$. You find $a_2 = 2$ and $b_2 = 0$ with
>
> $$a_1/10 + a_2/100 \;<\; 1/8 \;\le\; 1/10 + 3/100 \qquad \text{and} \qquad b_1/10 + b_2/100 \;<\; \sqrt{2}/2 \;\le\; 7/10 + 1/100.$$
>
> Continuing that way, we find two sequences (a_i) and (b_i) of digits (0 to 9) such that
>
> $$a_1/10 + a_2/100 + \ldots + a_n/10^n \;<\; 1/8 \;\le\; a_1/10 + a_2/100 + \ldots + (a_n + 1)/10^n,$$
> $$b_1/10 + b_2/100 + \ldots + b_n/10^n \;<\sqrt{2}/2 \;\le\; b_1/10 + b_2/100 + \ldots + (b_n + 1)/10^n.$$
>
> By the definition of series convergence, the two series
>
> $$a_1/10 + a_2/100 + \ldots \qquad \text{and} \qquad b_1/10 + b_2/100 + \ldots$$
>
> *equal* 1/8 and √2/2 respectively. We abbreviate the series by the symbols (**decimals**) $.a_1a_2\ldots$ and $.b_1b_2\ldots$. It is in this sense that
>
> $$1/8 \;=\; .1249999\ldots \qquad \text{and} \qquad \sqrt{2}/2 \;=\; .707[\text{non-repeating sequence of digits}].$$
>
> Every member of (0, 1) is given by a decimal.
>
> Given how we chose it, the decimal for such a real is unique. Had we written "smaller than or equal to" in place of "smaller than," then we would have arrived at
>
> $$1/8 \;=\; .1250000\ldots.$$
>
> It is not hard to show that this example gives the only way different decimals can have equal values:
>
> $$.1250000\ldots \;=\; .1249999\ldots,$$
>
> in which one decimal ends in an unending string of 0's, the other in a string of 9's; the former's last non-0 is 1 greater than the latter's last non-9; and the value is a **decimal fraction** $m/10^n$, like
>
> $$1/8 \;=\; 125/1000,$$
>
> whose reduced denominator necessarily has no prime factors other than 2 and 5.

Cantor's argument for the uncountability of **R** was so ingenious that it became a technique and acquired a name, **Cantor's diagonal argument**.

> Look at any natural-numbered list of reals written as decimals. Write it as
>
> $r_1 \quad = \quad .c_1c_2c_3\ldots$
> $r_2 \quad = \quad .d_1d_2d_3\ldots$
> $r_3 \quad = \quad .e_1e_2e_3\ldots$
> $\quad\quad \ldots.$
>
> The real number
>
> $r \quad = \quad .(c_1 + 5 \bmod 10)(d_2 + 5 \bmod 10)(e_3 + 5 \bmod 10)\ldots$
>
> (signifying succession of decimal digits, not multiplication) is not on the list. This r is not r_1. The first digits of r and r_1 differ by 5, and we just stated that two such discrepant decimals cannot have the same value. Similarly r is different from r_2, r_3, ….
>
> We conclude that no list can name all the reals. It follows that no one-to-one function maps **N** *onto* **R**. The set of real numbers is **uncountable**.

(ii) power sets

Casting about for other uncountable sets, we come upon the set of subsets of **N**. For any set S, the class of its subsets is called the **power set** of S, symbolized by $\mathcal{P}(S)$. Cantor showed that for every set, the power set has more members than the set.

It is worthwhile doing **N** separately, just to revisit the diagonal argument. Suppose

$S_1, S_2, S_3, \ldots$

is a list of subsets of **N**. Define a subset S by putting into it:

1 iff 1 is not in S_1,
2 iff 2 is not in S_2,
3 iff 3 is not in S_3, ….

Then S disagrees with every subset on the list. Hence $\mathcal{P}(\mathbf{N})$ is unlistable. (Compare Exercise IX.C.2c:4.)

> For a general set T, let f be a one-to-one function from T to $\mathcal{P}(T)$. (There certainly *are* such functions:
>
> $t \to \{t\} \qquad$ for all $t \in T$
>
> describes one. Consequently $\mathcal{P}(T)$ is as least as numerous as T.) Let U be the subset of T consisting of those members t that are not elements of their images $f(t)$. Then U cannot be an image under f. If there existed u in T such that $U = f(u)$, then we would have a contradiction:
>
> u would be a member of U exactly if u is not a member of $f(u) = U$.
>
> Therefore f cannot be onto. We conclude that $\mathcal{P}(T)$ has greater cardinality than T.

In view of this result, there is a whole scale of increasingly bigger infinities. Write "$\langle$" to mean *has smaller cardinality than*. Then

$$\mathbf{N} \quad \langle \quad \mathbf{R} \quad \langle \quad \mathcal{P}(\mathbf{R}) \quad \langle \quad \mathcal{P}(\mathcal{P}(\mathbf{R})) \quad \langle \quad \ldots.$$

Cantor saw the associated paradox in his set theory. The theory did not rule out a set S "of all sets." This set would have the biggest possible cardinality, because its members would include every set's singletons. But $\mathcal{P}(S)$ would necessarily have bigger cardinality.

(iii) the two basic sets

We now show that the reals and the subsets of **N** are equally numerous.

First, match each subset of T of **N** with a sequence $(a_1, a_2, \ldots)$ of 0's and 1's, according to the rule

$a_k \quad = \quad 1 \quad$ if k is in T,
$\qquad\quad\;\; 0 \quad$ if k is not in T.

For example, the set of even naturals goes with (0, 1, 0, 1, 0, 1, …). Decide for yourself why the result is a correspondence between $\mathcal{P}(\mathbf{N})$ and the set $\mathcal{2}$ of all such 0-1 sequences.

Next observe that just as any real in (0, 1) is given by a decimal

$$.b_1b_2b_3\ldots \;=\; b_1/10 + b_2/10^2 + b_3/10^3 + \ldots,$$

so it is given also by a **binimal**

$$\$c_1c_2c_3\ldots \;=\; c_1/2 + c_2/2^2 + c_3/2^3 + \ldots.$$

[There is no such mathematical word, so we invent "binimal"?] You find the binimal, in which the permissible digits c_1, c_2, … are all 0 or 1, by adapting the "biggest fit" algorithm that worked for 10. A given real's binimal is unique, except for the **binary fractions** $m/2^n$. For those, say 1/8,

$$\begin{aligned} 1/2^3 \quad &= \quad \$0010000\ldots \\ &= \quad \$0001111\ldots \quad = \quad 1/2^4 + 1/2^5 + \ldots. \end{aligned}$$

Using the binimals that do not end in all 0's, we match

$$\$c_1c_2c_3\ldots \quad \leftrightarrow \quad (c_1, c_2, c_3, \ldots).$$

That defines a function from the interval (0, 1) to $\mathcal{2}$. The function is clearly one-to-one.

The members of $\mathcal{2}$ that are not images under the function are those of the form

$$(d_1, \ldots, d_k, 0, 0, 0, 0, \ldots),$$

plus (1, 1, 1, …). That subset of $\mathcal{2}$ is countable (Exercise 1). We have found a one-to-one function from (0, 1) to $\mathcal{2}$ that covers all but a countable subset of $\mathcal{2}$. It follows (Exercise 2) that there exists a one-to-one function mapping the interval *onto* $\mathcal{2}$. Therefore (using "≡" for *is equivalent to*)

$$\mathbf{R} \quad \equiv \quad (0, 1) \quad \equiv \quad \mathcal{2} \quad \equiv \quad \mathcal{P}(\mathbf{N}).$$

The finite set {1, 2, …, n} has cardinality 2^n (Exercise 3). We have symbolized the cardinality of **N** by $\aleph_0$. Accordingly, it is usual to symbolize the cardinality of $\mathcal{P}(\mathbf{N}) \equiv \mathbf{R}$ by $2^{\aleph_0}$, which is in turn sometimes called $\aleph_1$. (Then $\mathcal{P}(\mathbf{R})$ has $\aleph_2 = 2^{\aleph_1}$ members, $\mathcal{P}(\mathcal{P}(\mathbf{R}))$ has $\aleph_3 = 2^{\aleph_2}$, ….)

We now know that **N** is less numerous than **R**. Cantor could not find any intermediate cardinality. He conjectured that none exists. The conjecture was called the **continuum hypothesis**. (More broadly, the **generalized continuum hypothesis** says that if S is infinite, then $\mathcal{P}(S)$ has the next biggest cardinality.) It was one of the deepest questions mathematics ever encountered. The answer took a remarkable form and required half a century to decide.

Exercises IX.C.2d

1. Prove that the sequences of 0's and 1's with just finitely many 1's (so that they end in all 0's) constitute a countable subset of $\mathcal{2}$. (Hint: Either the real numbers given by the corresponding binimals or Exercise IX.C.2c:4 can lead to the proof.)
2. Assume S is infinite. Suppose f maps S one-to-one into T, such that the images under f make up all but the countable (maybe finite) subset $\{t_1, t_2, \ldots\}$ of T. Show that there exists a correspondence between S and T.
3. Show why the power set of {1, 2, …, n} has 2^n members. (By Cantor's argument, $2^n > n$. Does that agree with our previous comparison?)
4. Show that there is a correspondence between **R** and the set **C** of complex numbers. (Hint: The class of subsets of **N** that hold only even numbers, its complement (the class of subsets that hold some odds), and the full $\mathcal{P}(\mathbf{N})$ are all equally numerous.)
5. Which is more numerous, the set of algebraic numbers, or the set of transcendentals?

Section IX.D. The Oldest Deductive System

In geometry, the first third of the nineteenth century brought a revolution that took most of the rest of the century to gain understanding and acceptance.

1. Non-Euclidean Geometry

a) Bolyai and Lobachevsky

Lambert's lament notwithstanding, attempts to prove the parallel postulate continued. The next noteworthy try came independently and almost simultaneously from a Russian academic and a Hungarian cavalry officer. Nikolai Ivanovich Lobachevsky (Lovachevskii?, 1792-1856) and János [YAH-nosh] Bolyai (1802-1860) made their discoveries in the 1820's (Bolyai first), then published them in 1829 and 1832 (Lobachevsky first). They denied the parallel postulate, in Playfair's form, and sought to find a contradiction.

The Bolyai-Lobachevsky Postulate. Given a point off a line, there exist *multiple* lines through the given point parallel to the given line.

(i) the Bolyais

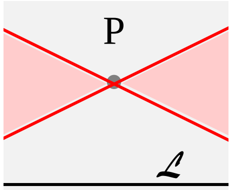
One of those working at the parallel postulate was Farkas Bolyai, a Hungarian teacher who had been a classmate of Gauss and kept in touch with him. Farkas tried to steer János, his son, into physics. When János told his father that he had become obsessed with the postulate, Farkas begged him to abandon it; see Boyer. János instead went on to build an unimpeachably logical geometry built on the Bolyai-Lobachevsky (hereafter "multi-parallel") postulate. He expanded the theory into a tract called *The Absolute Science of Space*. It must have been impossible to find a publisher for such revolutionary ideas, because the document finally appeared as an appendix to his old man's textbook of 1832.

(ii) Lobachevsky

Lobachevsky became professor at Kazan (400? miles east of Moscow). He saw how contrary to accepted wisdom his ideas were, and called them "imaginary geometry." Maybe he need not have worried: Published in Russia, his work was at first practically unnoticed in the French-German mathematical world.

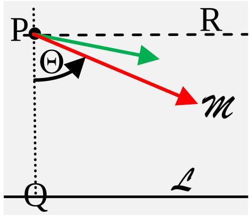
(iii) some theorems

Under the multi-parallel postulate, the geometry of Saccheri applies. (See section VIII.A.1.) Thus, the angles in a triangle sum to less than a straight angle, and the shortfall below a straight angle measures the area. In a Saccheri quadrilateral, the base and summit diverge to either side of the median, the two sides exceed the median, and the summit angles are acute.



The postulate has some striking consequences. [There is a great place to see them online, Henry Parker Manning's 1901 *Non-Euclidean Geometry* at Project Gutenberg.] For example, if P is off line ℒ, then there are two lines (red in the figure at left) that are the limiting positions for parallels to ℒ. That is, the two lines and any line through P into the (pink) zone between them are parallel to ℒ, and any line through P going outside the zone necessarily meets ℒ.

At right, we have point P off line $\mathcal{L}$, the perpendicular (dotted) from P meeting $\mathcal{L}$ at Q, and the perpendicular PR (dashed) to PQ at P. Line PR is parallel to $\mathcal{L}$, because it makes congruent a/i angles. By the postulate, there must be other parallels (green) to $\mathcal{L}$ through P. Look at just those parallels that go down to the right, making acute angles to the right of segment PQ at P; analogous statements hold for the ones going down to the left.



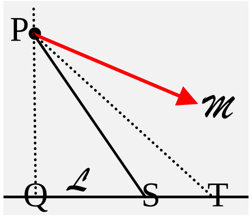
Those acute angles are all positive. Hence their measures have (what we now call) a greatest lower bound Θ. Let $\mathcal{M}$ denote the line (red in the figure) at that angle. Because Θ is a lower bound, any line making a smaller angle at P is not a parallel to $\mathcal{L}$. Its meeting with $\mathcal{L}$ has to be to the right, because to the left such a line stays above PR.

Any line making a bigger angle must be parallel to $\mathcal{L}$. Imagine that the green line makes an acute angle Θ + 1° with PQ. Because Θ is the GLB, that bigger angle is not a lower bound. Therefore there is a parallel $\mathcal{N}$ (not shown) that makes a smaller angle, say Θ + 0.2°. To the right, the green line stays above $\mathcal{N}$; to the left, it stays above PR; and $\mathcal{N}$ and PR both stay above $\mathcal{L}$. Therefore the green line does not meet $\mathcal{L}$.

As for $\mathcal{M}$ itself, take a line that intersects $\mathcal{L}$ at say S, as in the figure at right. Put T further right along $\mathcal{L}$. Angle QPT has to be less than or equal to Θ; if it exceeded Θ, then PT would not meet $\mathcal{L}$. Therefore

angle QPS < angle QPT ≤ Θ.

That says $\mathcal{M}$ is not line PS. We conclude that $\mathcal{M}$ does not meet $\mathcal{L}$ to the right.



It turns out that just as $\mathcal{M}$ is the limiting parallel to $\mathcal{L}$ at P, so $\mathcal{L}$ is the limiting parallel to $\mathcal{M}$ at Q. Indeed, at every point on either line, that line is the limiting parallel to the other. Lobachevsky referred to $\mathcal{M}$ as **the parallel** toward the right at P; for him, the lines above $\mathcal{M}$ (including PR) were simply "non-intersecting lines." Accordingly, he would have said that if $\mathcal{M}$ is parallel to $\mathcal{L}$, then $\mathcal{L}$ is parallel to $\mathcal{M}$; and that parallel lines are parallel at every one of their points. (We will stick to our usage: "Parallel" refers to lines in one plane that have no common point.)

Angle Θ is called the **angle of parallelism** for $\mathcal{L}$ at P. Considerations of symmetry show that Θ depends just on the distance from P to $\mathcal{L}$, not separately on the point and line (Exercise 1). Lobachevsky produced the formula

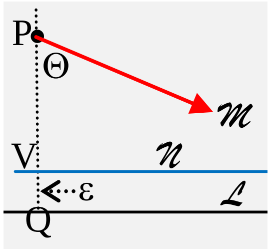
$$\Theta = 2 \tan^{-1} (e^{-[\text{length of PQ}]}).$$

(See its development starting at Harding's page 42.) That function decreases from limit π/2 when the length of PQ is almost 0 (and PR is practically the only non-intersecting line) toward limit 0 as the length of PQ tends to infinity (PQ is practically the only intersecting line).

Because Θ decreases with distance, we can infer that $\mathcal{M}$ is asymptotic to $\mathcal{L}$ in a peculiar way.

Draw the perpendicular (blue) $\mathcal{N}$ to PQ rightward at V, a small distance ε above Q. By the formula, the angle of parallelism for $\mathcal{N}$ at P is more than Θ. Therefore $\mathcal{M}$ crosses $\mathcal{N}$, no matter how small ε is.

That does not say the distance from $\mathcal{M}$ to $\mathcal{L}$ approaches zero. At the place where $\mathcal{M}$ crosses $\mathcal{N}$, the distance to $\mathcal{L}$ *has to be more than* ε. The perpendicular from that place to $\mathcal{L}$ is the right side of a Saccheri quadrilateral with median VQ; that side must exceed VQ.



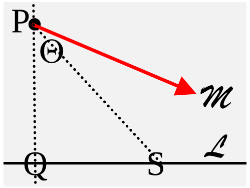
(iv) Gauss

No doubt worried that his son’s contrarian geometry would damage the boy’s reputation, Farkas Bolyai wrote for advice to his old friend Gauss. The latter acknowledged being impressed that János had been bright enough to rediscover what Gauss had thought up fifteen years before. (He modestly forbore to praise his own ideas.) It was not merely a catty remark: Gauss had always kept journals, and they clearly showed his own development of what he named “non-Euclidean geometry.”

None of the three—Gauss, Bolyai, or Lobachevsky—arrived at a contradiction. It is clear that by 1830, all three believed that Lambert’s suspicion was right, that the parallel postulate is independent of (not provable from) the others of Euclid. It is remarkable, both Struik (page 167) and Boyer observe, that three men separated by geography and culture should nearly simultaneously discover ideas that had eluded geometers for more than twenty centuries.

Exercises IX.D.1a

1. Show that if the distance from point P to line $\mathcal{L}$ equals the distance from point U to line $\mathcal{K}$, then the angle of parallelism for $\mathcal{L}$ at P equals that for $\mathcal{K}$ at U.
2. In the figure at right, S is moving rightward along $\mathcal{L}$. Show that as S goes out toward infinity, angle QPS approaches the angle Θ of parallelism for $\mathcal{L}$ at P.



b) Riemann

In 1854, Riemann had his *habilitation*, a sort of introductory presentation where a recent doctor shows off his capacity for discovery. Titled *On the Hypotheses That Lie at the Foundation of Geometry*, it was a watershed. Riemann introduced an overarching view of geometry based on the concept of “manifold,” an analytic generalization of space and of surface, not restricted to three dimensions. This global outlook makes it possible to view all models of geometry, Euclidean or not, as particular instances of more general structures.

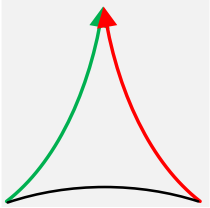
(i) different models

To illustrate “particular instances,” look at three of them.

Picture first an ant at one corner of a basketball court painted onto the middle of a large area of flat cement. As he sees it, he is standing on an endless plane. The lines delineating the court look perfectly straight. Opposite sides of the court are equidistant lines. Adjacent sides meet at right angles, so the boundaries form a quadrilateral with angle sum 360°. His is a Euclidean world.

Picture next an exceptionally tiny ant standing at the middle of an ordinary horse saddle. (The picture at right is from aliexpress.com.) The few square inches surrounding him look to him like part of a plane, in which Euclidean geometry reigns, at least approximately. It is necessarily that way in any geometry. At small scale, triangles have small area, have therefore angle sums indistinguishable from 180°.

We, standing back and able to see the big picture, know the saddle's surface curves. (Indeed, it curves in a strange way. Toward front or back, the path curves upward. To either side, the path curves down.) There are no Euclidean lines confined to the surface. Instead, the ant's "lines" are those curves that offer the shortest distance between points on the saddle, the geodesics. We can see that a geodesic from the left side toward the front curves to the left (green-arrowed curve in the figure at left), one from the right side curves right (red). The (black) connector between their starting points completes a triangle whose angles sum to less than 180°.



The saddle's surface is called a parabolic hyperboloid. Because the geometry on it has the features studied by Saccheri, Bolyai, and Lobachevsky, that kind of geometry is called "hyperbolic."

Last, picture a creature confined to a sphere, like an ant standing on a globe—or a pre-ballooning human standing on a slightly flattened sphere 8000 miles across. As we observed in section VIII.A.2, on a sphere the "straight" paths are great circles. Their geometry is considerably different from Euclid's. There are no parallels; indeed, any two distinct great circles meet at two points. From a given point off a "line," there are at least two perpendiculars to the line. That means there are triangles with angle sums of nearly 360°. From the association with spheres and ellipsoids, the geometry is called "elliptic."

(ii) one elliptic model

The standard example of Riemannian geometry avoids the sphere's problem—"lines" having two intersections—by considering diametrically opposite points to be identical. Equivalently, view only Earth's *northern* hemisphere with the understanding that a point on the Equator and the point halfway around the Equator from it are one and the same. [Save yourself a headache: Refer to a globe.] Thus, if you fly due south from New York along the 74° West meridian of longitude, some 2500 miles gets you to Bogotá, Colombia. After another 400 miles, you arrive at the Equator and instantaneously appear at the 106° East meridian. [Think of those video games wherein you can flee the pursuer by disappearing into the right side of the screen, rematerializing on the left.] From that spot (150 miles southeast of Singapore) you fly 6300 miles north along 106°E to the North Pole. Then you continue on the same heading, 3300 miles due south along 74°W back to New York.

In this example, all the lines have the same length,

$$2500 + 400 + 6300 + 3300 \text{ miles} = \text{half the sphere's circumference.}$$

Equality of lengths is not always the case; think of an ellipsoid, which in fact Earth resembles. However, always all the lines have finite length. That means that if you go far enough along a line without turning around, you return to where you started. By the same token, no line separates the "plane" into sides. In the Euclidean plane, each line separates the plane into two "sides," from either of which you have to cross the line to get into the other. On the Riemannian hemisphere, given a great circle going roughly east through Miami, you can avoid crossing it on your trip from New York to Bogotá by flying the 10,000-mile North Pole-Singapore-Equator route.

c) Klein

During his long life, including twenty-seven years at Göttingen, Felix Klein (1849-1925) unified much of the mathematics we have ascribed to the nineteenth century. He helped achieve wide understanding of Riemann's synthesis of analysis and geometry, and brought group theory into it.

(i) invariants under transformations

In what amounted to an inaugural address at Erlangen in 1872, Klein introduced the idea (the *Erlangen Programm*) of classifying geometries by means of the features that persist (**invariants**) under groups of transformations.

The transformation in the coordinate plane that assigns

$(x, y) \rightarrow (x + a, y + b)$, a and b fixed,

is a **translation**. For a fixed angle θ,

$(x, y) \rightarrow (x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta)$

defines a **rotation**. A **reflection** is given by

$(x, y) \rightarrow (-x, y)$.

The combinations of all those (**rigid**) transformations form a group under composition.

Each transformation in the group maps any line segment into another of equal length. By preserving length, it also transforms any triangle into a congruent copy (by SSS). Therefore it preserves areas, angles, parallelism—all the elements of Euclidean geometry. It might not preserve all the elements of Lobachevskian geometry: A reflection reverses the direction in which some pairs of lines are (Lobachevskian) parallels, *and not others*.

Finally, the group comprises *all* the transformations that preserve Euclidean relationships. That is, for any such preserving transformation, you can name a reflection, rotation, and translation (all: if necessary) that combine to give it. This group *characterizes* Euclidean geometry in the plane.

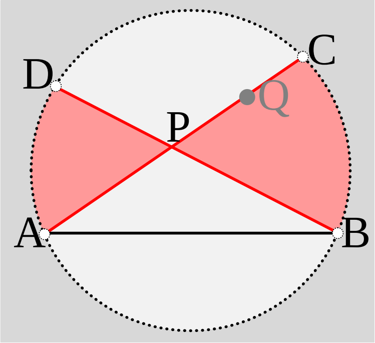
By exhibiting Bolyai-Lobachevsky properties and Riemannian properties as what remains invariant under other groups of transformations, Klein ended any controversy about the validity of hyperbolic and elliptic geometry.

(ii) one hyperbolic model

There is a simple picture by which Klein set up a model of Bolyai-Lobachevsky geometry within the Euclidean plane. Take the interior of some circle to act as (what we will call) the "K-plane" of points. For "K-line," take the part interior to the circle of any ordinary line that crosses the circle. Thus, a K-line is a chord of the circle, but with its endpoints removed.

Look first at relations among K-lines and points.

At right, we draw the circle dotted, to indicate that its points are not in the K-plane. The chord AB, minus its two ends (small white dots), constitutes a K-line <AB>. [The notation is not standard, but it does suggest endlessness.] Point P is not on <AB>. Clearly there are many K-lines through P K-parallel to <AB>, meaning that they have no points in common with (Euclidean) line AB *in the interior of the circle*. Draw the (red) extensions of AP past P to C on the circle and of BP past P to D on the circle. Those chords locate the K-lines <AC> and<BD> that are the leftward and rightward limiting K-parallels to <AB> through P.



You have to give distance a special definition to avoid having lines of finite length.

Imagine point Q (grey in the figure) on the segment PC. Klein defined the K-distance (our name) from P to Q, or the K-length l(PQ) of the K-segment PQ, by

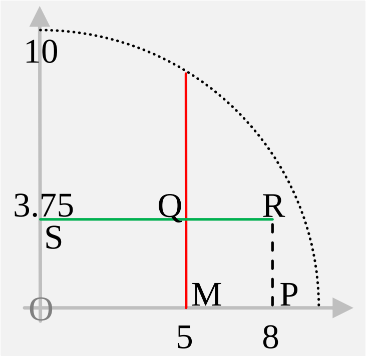
$l(\mathrm{PQ}) = \log_e (|\mathrm{PC}|\,|\mathrm{QA}| \,/\, |\mathrm{PA}|\,|\mathrm{QC}|)$.

(There, |segment| means the Euclidean length of the segment.)

Observe that the definition gives K-lines infinite K-length: As Q → C (or P → A) along <AC>, one factor in the denominator approaches zero, the two factors in the numerator approach positive limits, and $l(\mathrm{PQ}) \rightarrow \infty$.

With a changed definition of length, you have to accept that circles—defined as usual by equidistance from a point—get deformed toward the center of the overlying circle. You then have to define angle measure according to those other circles; you have to characterize the measure in terms of length. To illustrate the needed adjustments, consider characterizing perpendicularity.

> Use the coordinate-plane circle (figure below right) whose interior is given by
> $x^2 + y^2 \; < \; 100.$
> Put P at (8, 0). The K-line $x = 5$ (red) is the K-perpendicular bisector of K-segment OP. The K-perpendicularity is clear from symmetry, but remember that we want to characterize it by means of K-distance. The characterization is that $x = 5$ is the locus of points K-equidistant from O and P.
>
> The K-segment OP has K-length
> $l(\mathrm{OP}) \;=\; \log_e (\, 10[10 + 8] \,/\, 10[10 - 8] \,) \;=\; \log_e 9.$
> The place M(m, 0) halfway to P satisfies
> $\log_e (\, 10[10 + m] \,/\, 10[10 - m] \,) \;=\; 1/2 \log_e 9 \;=\; \log_e 3.$
> That forces $m = 5$; (5, 0) is the K-midpoint of OP. (See Exercise 1.)
>
> To illustrate that points along $x = 5$ are K-equidistant from O and P, Q(5, 3.75) is convenient, because it is on the radius ending at (8, 6). [In this model it is helpful to work with chords having integer coordinates at the ends, but it is *essential* that the coordinates be rational.] We have
> $l(\mathrm{OQ}) \;=\; \log_e (\, 10[10 + \sqrt{(5^2 + 3.75^2)}] \,/\, 10[10 - \sqrt{(5^2 + 3.75^2)}] \,)$
> $\;=\; \log_e (13/3).$ (Check the fractions.)
> (Notice that by Euclidean similarity, we can use x-differences instead of lengths:
> $l(\mathrm{OQ}) \;=\; \log_e (\, 8[8 + 5] \,/\, 8[8 - 5] \,) \;=\; \log_e (13/3).$)
> Points P and Q are on the chord ending at (0, 10) and (400/41, -90/41) (Exercise 2). Using the x-differences, we get
> $l(\mathrm{PQ}) \;=\; \log_e (\, [400/41 - 5][8 - 0] \,/\, [5 - 0][400/41 - 8] \,)$
> $\;=\; \log_e (13/3).$ (Check.)
> Thus, Q is K-equidistant from O and P. The same holds for all points on K-line $x = 5$.
>
> Finally, put R at (8, 3.75), S at (0, 3.75) (ends of the green line). Since PR is on the vertical chord ending at (8, 6), the y-differences give us
> $l(\mathrm{PR}) \;=\; \log_e (\, 6[6 + 3.75] \,/\, 6[6 - 3.75] \,) \;=\; \log_e (13/3).$
> Because PR is K-congruent to PQ, $x = 8$ is not K-perpendicular to $y = 3.75$. Instead, the K-perpendicular from P to SR goes to the K-midpoint of QR. [If you want to see how hard it is to work with K-lines along chords ending at irrational coordinates, try to calculate that K-midpoint.]



Numerous features of hyperbolic geometry show up in the figure. Triangle PQR is K-isosceles, so the base angles PQR and PRQ must be K-congruent K-acute angles. Quadrilateral OPRS is a Lambert (section VIII.A.2), with K-right angles at O, P, and S; it has to have a K-acute angle at R. Last, both pairs of K-lines having a common perpendicular diverge away from the perpendicular.

> RP and SO are both perpendicular to the x-axis, and they get further apart as you go up: OP has length $\log_e 9$, whereas $l(\mathrm{SR})$ is
> $\log_e (\, [\sqrt{(10^2 - 3.75^2)}][\sqrt{(10^2 - 3.75^2)} + 8] \,/\, [\sqrt{(10^2 - 3.75^2)}][\sqrt{(10^2 - 3.75^2)} - 8] \,) \;\cong\; \log_e 13.6.$
> Similarly, RS and PO are both perpendicular to the y-axis, and they diverge to the right:
> $l(\mathrm{OS}) \;=\; \log_e (\, 10[10 + 3.75] \,/\, 10[10 - 3.75] \,) \;=\; \log_e 2.2,$
> $l(\mathrm{MQ}) \;=\; \log_e (\, [\sqrt{75}][\sqrt{75} + 3.75] \,/\, [\sqrt{75}][\sqrt{75} - 3.75] \,) \;\approx\; \log_e 2.53,$
> $l(\mathrm{PR}) \;=\; \log_e (13/3).$

Exercises IX.D.1c

1. Show in our circle that the K-distance from (5, 0) to (8, 0) matches the K-distance (namely $\log_e 3$) from (0, 0) to (5, 0)
2. a) Show that (5, 3.75) and (8, 0) are on the line given by
   $y = -5/4\ (x - 8)$.
   b) Show that the line in (a) and the circle $x^2 + y^2 = 100$ intersect at $x = 0$ and $x = 400/41$.
3. Use a scientific calculator to check that in right triangle OMQ,
   $l(\text{OM})^2 + l(\text{MQ})^2 \approx (\log_e 3)^2 + (\log_e 2.53)^2$ is less than $l(\text{OQ})^2 = (\log_e 13/3)^2$.
   The Pythagorean theorem is true in, and only in, Euclidean geometry.

(iii) the implication

The picture gives a model with all the features of Bolyai-Lobachevsky geometry. The fact that the model is built out of Euclidean parts yields a remarkable inference.

Recall what Bolyai and Lobachevsky had in mind: Denying the parallel postulate for the purpose of using Euclid's other postulates to reach a contradiction. If such a contradiction exists, then it is to be found in Klein's picture. But Klein's picture is assembled from Euclidean elements. If the multi-parallel postulate leads to a contradiction, then that contradiction exists in Euclidean geometry.

Klein's model establishes **relative consistency**. It does not guarantee that Bolyai-Lobachevsky geometry is a consistent deductive system. What it shows is that if Bolyai-Lobachevsky geometry is inconsistent, then so is Euclidean geometry. To look at it a different way, Klein demonstrated that non-Euclidean geometry is as defensible and useful, as either a deductive system or a description of the world, as what the intellectual heirs of the Greeks had revered for two millennia.

2. The Way of the World

In the face of the non-Euclidean geometries, it was still possible to believe that Euclid gave, if not the only valid system, at least the true picture of the geometry of the universe. Bolyai did not share that opinion. He thought his geometry was as likely as Euclid's to describe the actual universe at large scale. He made an interesting prediction: The determination of which system was the right description would come from science, not mathematics—from experiment, not deduction. He was mostly right. The experiment came in 1919, and it left no doubt that a Riemannian geometry gives the most accurate description of the universe.

a) Maxwell's equations

We go first to the Scotsman James Clerk [pronounced "Clark"] Maxwell (1831-1879). Maxwell and some fellow Brits were great contributors to mid-nineteenth-century algebra and partial differential equations. (Look up William Hamilton, George Green and Lord Kelvin, and George Stokes.) His contributions to mathematical physics were out of this world.

By 1859, he had applied Lagrange's mechanics to show that the rings of Saturn cannot be solid. A solid ring, he proved, would be unstable. It would be torn apart by gravity, which would send pieces falling to the planet or escaping to space. Instead the rings must be thin bands—small in thickness, compared to width and diameter—of rocks, ice balls, and other such loose debris orbiting the big planet. (Modern times produced evidence from the two Voyager spacecraft, travelling roughly 1977-1981. Now we have confirmation, plus many new mysteries of orbital mechanics, from the Cassini probe NASA put into Saturn orbit in 2004. "Cassini" honored Giovanni Domenico [later Jean Dominique] Cassini, who visually discovered that the seeming "ring" of Saturn has a gap separating it into, by his count, two rings.)

In 1867, Maxwell adapted Gauss's normal distribution to the energies of particles in an enclosed gas. The result related properties of gas molecules to properties of the gas as a whole (meaning global, average properties we can measure, like pressure and temperature. Look up the ideal gas law.)

His crowning achievement was Maxwell's equations, published in 1873. They are four partial differential equations that completely describe electricity, magnetism, and the relation between the two. From around 1825, aspects of electricity, magnetism, and their connection had been studied by, among others, the Frenchman André-Marie Ampère, the Germans Gauss and his (physics) collaborator Wilhelm Weber, and the great English experimentalist Michael Faraday. All their discoveries—about charges and electric fields, charge flow (current) creating magnetic fields, and magnetic fields inducing current and electric fields—were subsumed into Maxwell's four PDE's. (Here again Maxwell synthesized local, micro-scale phenomena with global, large-scale. Differential equations are necessarily local descriptions. Ampère, Gauss, and Faraday had put their laws in global forms.)

In addition, the equations implied that a variable current would create an electromagnetic signal. The signal would consist of a joint vibration, a wave carrying an electric field of sinusoidally varying strength, moving together with (but at right angles to) a similarly variable magnetic field. Those "radio waves" were undetectable until Heinrich Hertz (Wikipedia®) established their existence in the 1880's.

According to the equations, the speed of propagation of such waves depends on two parameters that you can measure in the laboratory. Roughly speaking, one ("permeability") is the resistance of space to electric fields, the other ("permissivity") the receptivity (opposite of resistance) to magnetic fields. From the measured values, the calculated speed of electromagnetic waves is precisely the speed of light. The conclusion was inescapable: Light is simply one form, covering a tiny range of wavelengths, of electromagnetic radiation. As such, light is in some way fundamentally connected to space.

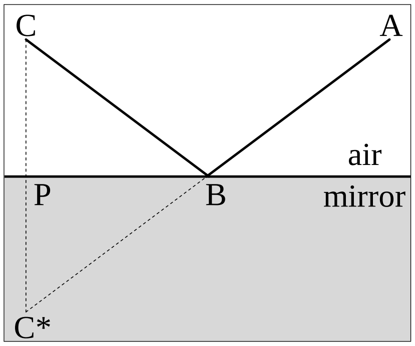
b) Fermat's principle

That name is attached to a minimization principle that explains all that was known in 1662 about the behavior of light. (Check Wikipedia®, which says that the concept had already appeared in the immortal ibn al-Haytham's *Book of Optics* (mentioned in section V.A.4a).)

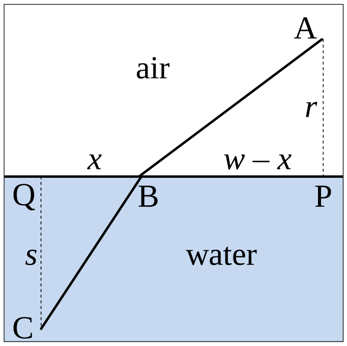
Fermat's Principle. Light travels from point to point by the path that requires the least time.

Apply the principle to reflection first. In the figure below right, we have a ray of light from point A in the air, travelling to C after reflecting from a mirror at unknown point B. All the route is in air, at constant speed. Therefore the path of least time is the path of least distance from A to the mirror to C.

> Let the perpendicular from C reach the mirror at P, and let C* be the mirror image of C, located on the extension of CP by an equal length. Wherever B is, the right triangles CPB and C*PB are congruent by SAS. Therefore the broken line ABC is as long as ABC*; the shortest length ABC goes with the shortest ABC*. We know the latter occurs when ABC* is a straight line. In that case, the acute angle between AB and the mirror is vertical to angle C*BP, which is congruent to angle CBP. To take the least time, light travels so as to make the incoming angle with the mirror equal to the outgoing.



To explain refraction, we have light travelling from point A in one medium, call it air, to C in a different medium, say water. Clearly the fastest path is a straight segment AB to B on the surface, then a second segment BC. Putting B on the line AC gives the smallest distance. However, if the speed *v* of light in water is smaller than the speed *V* in air, then the drawn path, with a shorter water segment than the straight, might afford a reduced time. Let us find where B has to be.



First, there really is a path of least time. Let AP be the perpendicular from A to the surface, CQ the perpendicular from C. Write the constant lengths as AP = *r*, CQ = *s*, QP = *w*. At speeds *V* and *v*, the travel time is

AB/*V* + BC/*v*.

Assume B is *x* (possibly negative) to the right of Q. Then the time is

$$t(x) \;=\; \sqrt{r^2+(w-x)^2}/V + \sqrt{s^2+x^2}/v.$$

You can check that $t(x) > t(0)$ if $x < 0$ and $t(x) > t(w)$ if $x > w$, but think of it more simply. If B is leftward of Q, then the air segment AB exceeds AQ and the water segment BC exceeds QC; the path has to take more time than AQC. Similarly, if B is rightward of P, then the path takes longer than APC. Hence the least time, if there is one, is to be found for $0 \le x \le w$. We know that for that interval of *x*, the continuous function $t(x)$ *must* have a minimum value.

By Fermat's theorem [even if we aren't handling a polynomial], the minimum occurs where the derivative of *t* is zero. We find the derivative by Fermat's method. [What else?] We have

$$t(x+h) - t(x) \;=\; \sqrt{r^2+(w-[x+h])^2}/V + \sqrt{s^2+[x+h]^2}/v - \sqrt{r^2+(w-x)^2}/V - \sqrt{s^2+x^2}/v.$$

Combining by like denominators and rationalizing the numerators, we get

$$t(x+h) - t(x) \;=\; \frac{[r^2+(w-[x+h])^2]-[r^2+(w-x)^2]}{V\left(\sqrt{r^2+(w-[x+h])^2}+\sqrt{r^2+(w-x)^2}\right)} + \frac{[s^2+[x+h]^2]-[s^2+x^2]}{v\left(\sqrt{s^2+[x+h]^2}+\sqrt{s^2+x^2}\right)}.$$

Simplifying and dividing by *h*, we get

$$\frac{t(x+h)-t(x)}{h} \;=\; \frac{-2w+2x+h}{V\left(\sqrt{r^2+(w-[x+h])^2}+\sqrt{r^2+(w-x)^2}\right)} + \frac{2x+h}{v\left(\sqrt{s^2+[x+h]^2}+\sqrt{s^2+x^2}\right)}.$$

We set $h = 0$ on the right, and we reach

$$t'(x) \;=\; \frac{x-w}{V\sqrt{r^2+(w-x)^2}} + \frac{x}{v\sqrt{s^2+x^2}}.$$

That derivative is zero if

$$\frac{w-x}{V\sqrt{r^2+(w-x)^2}} \;=\; \frac{x}{v\sqrt{s^2+x^2}}.$$

Look at the picture: That last line says

$$\cos(\text{angle ABP})/V \;=\; \cos(\text{angle CBQ})/v.$$

That is the refraction law; check it against the (literally) normal version in Exercise V.A.4:1.

c) Einstein’s theories

In 1905, Albert Einstein (1879-1955) published four papers. One was on the size of molecules, along the lines of some work (1815 and after) by the great chemist Amedeo Avogadro. The next one, building on the previous, explained Brownian motion. That would already have made science remember Einstein’s name. A third paper, actually written first, explained the photoelectric effect (which had been discovered by Hertz). That one received the Nobel Prize in physics for 1921. (It took that long for science to verify—and more fundamentally, understand—Einstein’s explanation. The paper was also a triumph for Newton’s particle theory of light, four decades after Maxwell had staged the triumph for Huygens’s wave theory. Ever since, much of physics has needed to account for “wave-particle duality.”) Finally, Einstein wrote the big one.

[Our next must-read for anybody interested in science and its history is *Annus Mirabilis*, by John and Mary Gribbin. It gives history and description for Einstein’s discoveries of that year, including how a letter Einstein wrote *later* noted that the fourth paper implies the most famous equation of all time.]

(i) the special theory

The fourth paper introduced the special (same as “restricted,” to unaccelerated motion) theory of relativity. This was a revolution: It established that Newton’s equations fail at the atomic scale, where speeds comparable to the speed of light are possible.

> Einstein took Maxwell’s equations as basic properties of space, holding unchanged in every unaccelerated frame of reference. With that assumption comes the postulate that light has the same speed in every frame. The postulate implies that in a frame moving relative to us, mass, length, and time *are all changed* from their values in our frame.
>
> For mass, the change is that a mass we would measure at m_0 becomes
>
> $m = m_0/\surd(1 - v^2/c^2)$, c denoting “the” speed of light,
>
> in a frame we see moving at speed v. In Newtonian mechanics, mass is constant. If you do work on mass m_0, the added energy shows up in kinetic energy $m_0 v^2/2$. In relativistic mechanics, some of the added energy turns up as added mass. Indeed, as v increases toward c, the changed mass m approaches infinity; added energy goes more and more to added mass, not to increasing v.
>
> The changes in length (in the direction of motion) and time force a change in the way speeds combine. In 2006, a spacecraft named *New Horizons* left Earth at about
>
> $v_1 = 36000 \text{ mi/hr} = 10 \text{ mi/sec}.$
>
> That was the highest speed ever produced by human-made propulsion. Earth orbits the Sun at about
>
> $v_2 = 67000 \text{ mi/hr} \approx 19 \text{ mi/sec}.$
>
> What was the craft’s speed relative to the Sun? Newton would have combined the speeds as Galileo prescribed, by what physics calls the “Galilean transformation”:
>
> $v = v_1 + v_2,$
>
> about as fast as Mercury orbits the Sun. Einstein said you have to apply the “[Hendrik] Lorentz transformation”:
>
> $v = (v_1 + v_2)/(1 + v_1v_2/c^2).$
>
> Notice that the correction is by a factor of
>
> $(1 + v_1v_2/c^2)^{-1} \approx$ [binomial theorem] $1 - (10)(19)/(186000)^2 \approx 0.999\,999\,994;$
>
> even at speeds enormous by human standards, Newton is almost perfect. But if we were talking about a hydrogen nucleus moving at $V_1 = 2c/3$, somehow firing off its proton in the same direction at speed $V_2 = 3c/4$, then our reading of the combined speed would be
>
> $V = (2c/3 + 3c/4)/(1 + [6c^2/12]/c^2) = 17c/18.$
>
> That is still below c, and much different from the Newtonian value.

> Notice that if the nucleus, moving at speed V_1 away from us, emitted a *photon* back toward us at speed c relative to the nucleus, then we would measure the photon's (signed) speed as
> $$(V_1 + -c)/(1 + V_1[-c]/c^2) \ = \ c^2(V_1 - c)/(c^2 - V_1c) \ = \ -c.$$
> In both frames of reference, light has the same speed.

Thus, light—and all electromagnetic radiation—is tied up inextricably with both space and *time*.

(ii) the general theory and the geometry of space

In his last few years, the German mathematician Hermann Minkowski (1864-1909; visit St Andrews) showed that you could model the interrelated whole geometrically as a Riemannian four-dimensional continuum that he called **space-time**. In space-time, Fermat's principle suggests, light follows the geodesics. Those geodesics have to conform to the curvature of the Riemannian structure. Hence you can determine the geometry of space by checking the extent to which light follows curved paths instead of Euclidean lines. Einstein (originally unhappy with the idea of Minkowski's geometry's swallowing up Einstein's physics) later embraced the geometry and proposed a way to check.

Continuing work on relativity, Einstein by 1915 developed the general theory. Part of it explained gravity, not as force acting at a distance (a notion even Newton disliked), but as a warping of space-time. For an analogy, think of the Sun deforming its neighborhood in space-time into a bowl shape. The planets and other wanderers must follow closed orbits conforming to the curvature of the bowl, or fall to the bottom, or travel curved paths fast enough to escape over the edge. Einstein concluded that light is likewise confined to certain curved paths.

Newton, with his light "corpuscles," figured gravity should pull light. Einstein made certain with a thought experiment worthy of Galileo (section VI.D.3a(i)). Imagine, Einstein reasoned, being in an elevator cab way out in space, away from any sources of gravity. A force is tugging at the cable on the cab's roof, accelerating it at 32 ft/sec^2. You inside the cab feel your feet pressing against the floor. If you hold a ball in your hand, its inertia resists the acceleration; your hand has to provide the force to make it stay, as you see it, in place. If you release the ball, it continues moving up—as we outside the elevator see it—at constant speed, while you gain 32 ft/sec every second. Accordingly, it seems to you that the ball is "falling" exactly as it would have if the elevator had been stationary at the surface of Earth. You cannot distinguish between your elevator situation and what would happen in Earth's surface gravity.

Einstein formalized that inability to distinguish into the **equivalence principle**: There is no physical experiment you can do to decide whether you are stationary in a gravitational field or accelerating where there is no matter-induced gravity. (Notice, then, that the equivalence between gravitational field and accelerated frame puts the former under the general theory rather than the special.)

[For the combined-speed issues in (i), Einstein's examples always had a train, going say 50 mi/hr, with a passenger throwing a ball forward at 40 mi/hr, or throwing a light beam at c. Similarly, for accelerated frame of reference, he put a person on a lifting platform. Maybe he had not ridden elevators or cars. Our experience includes elevators with stomach-churning starts or stops. We have also enjoyed cars going too fast around curves, so that the hamburger we set on the next seat slides toward the outside of the curve, pulled by a "g-force" Huygens would put at

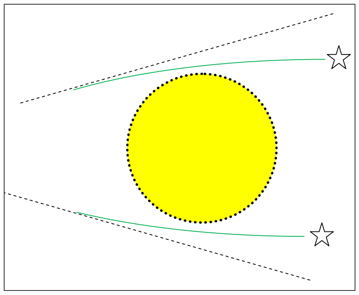
$$\text{(mass of burger)(speed of car)}^2\text{/(radius of curvature).}$$

We can visualize also the g-forces we know act on astronauts as they speed up at launch or slow down on reentry, events Einstein definitely missed.]

Now imagine that we who are outside the elevator set up a beam of light perpendicular to the cable. As the elevator passes the beam, a short burst of light enters the cab through a small hole in the side. We, looking into the cab, see the burst proceed across in the line of the beam, parallel to the (moving) planes of floor and ceiling. You, because the floor is accelerating toward the burst and the ceiling accelerating away, conclude that the light is accelerating downward. That is what the ball did, under the same gravity that is holding your feet to the floor. You conclude, by the equivalence principle, that light deflects under gravity, just as material objects do.

> Keep some perspective. Light travels about 10^9 ft/sec. If your elevator is 6 ft across, then the burst takes 6×10^{-9} sec to cross the cab. In that time, the burst falls
> $$at^2/2 \;=\; 32 \text{ ft/sec}^2 \,(6 \times 10^{-9} \text{ sec})^2\,/2 \;\approx\; 6 \times 10^{-16} \text{ ft.}$$

It takes a lot more gravity than Earth has to create measurable deflection . Even the Sun is somewhat weak, but for Einstein it had to suffice. He predicted that tight measurements of star positions near the Sun would demonstrate the effect: As the figure at right suggests, the curvature of the light paths (green) puts the apparent positions of the stars along the dashed lines, further from the Sun than their known positions imply. The opportunity to test came with the total eclipse of May 29, 1919. The Sun was within the vee-shaped group of stars ("the Hyades") that define the Bull's face, stars with precisely-known locations. The measured displacements were in sufficient agreement with Einstein's calculations to confirm his prediction. Bolyai's prophecy had come true: Scientific experiment decided what model best describes the geometry of space.



3. Revisiting Euclid

Even losing its place as descriptor of the universe, Euclidean geometry remains valuable. One factor in its favor is that it approximates practically any geometry at small-enough scale. (See Exercise 1. The mathematical term is that geometries are "locally Euclidean.") There is almost always a threshold below which Euclidean geometry is, like Newtonian mechanics if you stay slower and farther from the Sun than Mercury, good as gold.

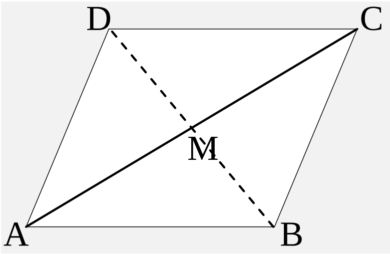
In the nineteenth century's spirit of axiomatization, we will look at some logical deficiencies in Euclid's system and a late-century deductive system that addressed them.

a) unstated assumptions

We noted in section III.A.5b that even contemporaries of Euclid objected to inferences that were intuitively undeniable but not justified by reference to axioms and theorems. We had an example back there. Let us view a similar one.

The proposition at hand is that the diagonals of a parallelogram bisect.

> At right is parallelogram ABCD. First draw diagonal AC. Triangles ABC and CDA are congruent by ASA. (Remember that we are back in Euclid-land: Parallels form congruent a/i angles.) Therefore the opposite sides are congruent.
>
> Now draw BD (dashed), intersecting AC at M. Triangles ABM and CDM are congruent by ASA. Therefore AM = CM and BM = DM. That says the diagonals bisect.



The proposition is true, but the argument strays out of the deductive system. It *assumes* that the *segments* AC and BD intersect within the parallelogram. (Their *lines* do have to meet.) As we observed earlier, words like "within" are not defined in Euclid. [It isn't just Euclid. We said that Lobachevsky's limiting parallel (section IX.D.1a(iii)) to AB through D would descend toward AB going, say, rightward, but would then have to stay "above" DC going leftward.]

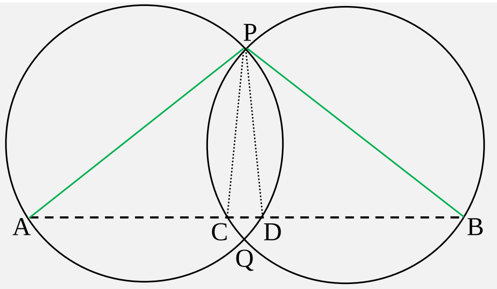
> Typically, it is possible to substitute unbiased arguments. Given parallelogram ABCD, draw only AC and let M be its midpoint. Draw the two segments BM and DM. Because opposite sides are congruent, triangles MAB and MCD are congruent by SAS. Therefore angles AMB and CMD are congruent. Since AMC is straight, angles AMB and CMB add up to a straight angle. Hence
>
> angle CMD + angle CMB = straight angle.
>
> In other words, DMB is one segment, the second diagonal. By the triangle congruence, DM = BM; M is the common midpoint of the diagonals.

Notice that the second argument depends on determining a midpoint and drawing two segments. Both are justifiable from the axioms and theorems.

b) paradoxes

The trouble with unstated assumptions—they are always suggested by pictures—is that they lend support to fallacious arguments, leading to seeming contradictions. We look at three well-known examples of such “proofs.”

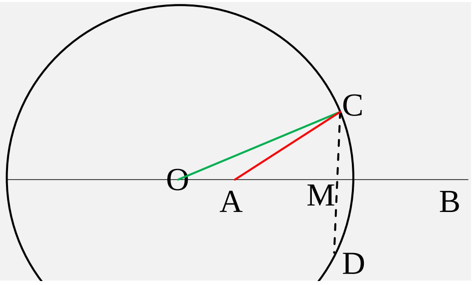
The first claims to build a triangle with two right angles.



> Let two circles intersect at P and Q, as in the figure at right. Draw the diameters (green) PA in one circle and PB in the other. Let AB intersect the two circles at C and D. Then angle PCB, being inscribed in a semicircle, is a right angle. Similarly, angle PDA is a right angle. Therefore triangle PCD has two right angles.

Clearly the figure must be ill drawn. However, you do have to make sure that the pictured situation is impossible; do Exercise 2.

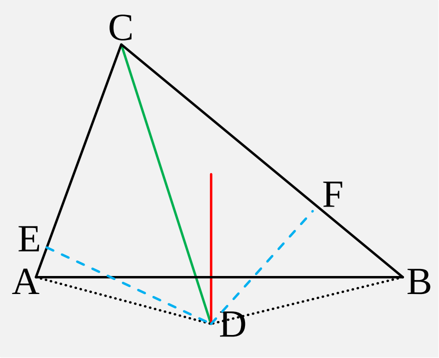
The next one purports to show that a point interior to a circle is actually on the circle.



> At right, we have a circle of radius 1 centered at O. Point A is at (positive) distance OA < 1 from O. Choose B on the extension of OA so that OB = 1/OA. Let M be the midpoint of AB, and let the perpendicular at M meet the circle at C and D.
>
> Using the Pythagorean theorem twice, we have
>
> $$\begin{aligned} AC^2 &= AM^2 + CM^2 \\ &= AM^2 + (OC^2 - OM^2) \\ &= 1^2 - (OM^2 - AM^2) \\ &= 1 - (OM - AM)(OM + AM). \end{aligned}$$
>
> The first factor (OM – AM) equals OA. Since AM = MB, the other factor
>
> $$OM + AM = OM + MB = OB.$$
>
> Therefore
>
> $$AC^2 = 1 - (OA)(OB) = 1 - 1,$$
>
> and A = C is actually *on* the circle.

Resolve the paradox with Exercise 3. The evidence against the picture is plentiful.

The last one shows that every triangle is isosceles.



> Triangle ABC has BC > AC. The bisector (green in the figure) of angle C cannot be parallel to the perpendicular bisector (red) of side AB (Exercise 4). Let them meet at D. Drop the perpendiculars (blue dashed) from D to AC and D to BC, ending respectively at E and F; and draw the segments DA and DB (dotted).
>
> Points on an angle bisector are equidistant from the sides of the angle. Hence DE and DF are congruent. On a segment's perpendicular bisector, points are equidistant from the endpoints. Hence DA and DB are congruent. By hypotenuse-leg, right triangles DEA and DFB are congruent, as are DEC and DFC. We then have lengths
>
> CE = CF and EA = FB.
>
> By addition, CA = CB. (Now do Exercise 5.)

Exercises IX.D.3

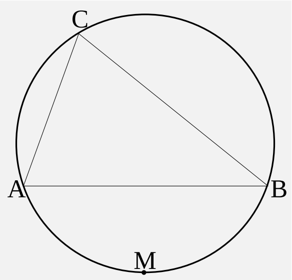
1. Refer to the non-Euclidean definitions in section IX.D.1c(ii). In Exercise 3 there, in right triangle OMQ the square $l(\mathrm{OQ})^2 \approx 2.15$ of the hypotenuse exceeds the sum

   $l(\mathrm{OM})^2 + l(\mathrm{MQ})^2 \approx 2.07$

   of the squares of the legs by about 4%. Calculate the corresponding excess for the right triangle with vertices at O(0, 0), T(0.4, 0), U(0.4, 0.3). [For your information: TU is on a chord whose ends are the points $(0.4, \pm\sqrt{[100 - 0.4^2]})$.]
2. In this subsection's second figure, with the two circles, show that Q *has to be* the place where AB crosses the circles. That dissolves "triangle PCD."
3. In the third figure, with B chosen to make (OA)(OB) = 1, show that the midpoint M of AB *has to be* outside the circle. That makes C, D, and the triangles disappear. (The problem amounts to showing that OA = $s < 1$ forces

   $[s + 1/s]/2 > 1.$

   You can do that any of three ways: by algebra, by calculus, or with a picture of squares.)
4. In the fourth figure, in which triangle ABC has BC > AC, show that the bisector of angle C cannot be perpendicular to side AB.
5. In the triangle figure, you might first suspect that D is improperly placed. It does, however, belong where shown, below the triangle. (If it were on AB or inside the triangle, the same argument would still appear to work.)

   a) At right, we have added the triangle's circumcircle. Arc AB—the arc that does not have C—is necessarily outside the triangle, and we see its midpoint M. Show that D = M: The bisector of angle C and the perpendicular bisector of side AB must meet at M.

   b) Given BC > AC, show that F (end of perpendicular from M to BC) has to be between B and C, whereas E (from M to AC) has to be *past* A along the line CA. That blocks the key addition at the end of the argument. (Hint: Use the circle to show that angle MBC has to be acute and angle MAC has to be its supplement.)

4. Hilbert's Fix

The last of the dons at Göttingen was David Hilbert (1862-1943). In 1900, he was the most influential mathematician in the world (despite the prestige of the Frenchman Henri Poincaré). Accordingly, for the 1900 International Congress of Mathematicians (compare 1936 in section III.A.7b(iii)), he was invited to deliver the keynote address.

At the suggestion of Minkowski, he chose to speak about unsolved problems. He picked his famous twenty-three "Mathematical Problems," not simply because they had stumped the math community, but because he judged that their solutions would engender new areas of mathematical research.

You can imagine how advanced they were, but there are three on which our material has touched. First on Hilbert's list was the continuum hypothesis, Cantor's guess that there is no cardinal strictly between $\aleph_0$ and $2^{\aleph_0}$. (No set is simultaneously more numerous than the natural numbers and less numerous than the reals.) Second was the consistency of arithmetic. Recall that Cantor's conception of set theory led to paradoxes. The second problem was the question of whether Peano's axioms might also lead to contradiction. The eighth problem was the Riemann hypothesis (section IX.B.6). [Consult Wikipedia® for Hilbert's list and more: ambiguity in some of the problems, inconclusiveness of some of the answers, and (just before the Summary) Hilbert's wonderful remark about the Riemann.]

Our interest is a list of postulates Hilbert assembled by 1899 for "incidence geometry," an axiomatization of Euclidean geometry. The list has about twenty axioms. ("About" because you can find listings with axioms combined, and others with single axioms split into multiple.) We will write them all, and draw inferences for many. For the others, we will only suggest what they imply.

a) incidence in one plane

A **planar** (**incidence**) **geometry** is an abstract structure, consisting of a set, together with a family of subsets of the set, satisfying three axioms.

Axiom I. Given two distinct members of the set, exactly one subset in the family has them.

Axiom II. Each subset in the family has at least two distinct members.

Axiom III. No subset in the family has all the members of the set.

We will immediately substitute the language of geometry for that of sets. The overlying set is **the plane**. Its members are **points**. The subsets in the family are **lines**. If a point is a member of a line, we say that **the point is on the line**, and **the line crosses the point**.

We recast the axioms in geometric language.

Axiom 1. Given two distinct points, there is one and only one line that crosses both. (Two distinct points determine a line.)

Axiom 2. Every line crosses at least two distinct points.

Axiom 3. No line crosses all the points.

[Using "distinct" all the time gets old fast. Let us agree to use "two," "three," and onward to *mean* what they normally do, two or more unequal things.

One of my students pointed out that these axioms are all universal statements. As such, they are vacuously true in the empty set. Geometry *in vacuo* is a boring pursuit. The easiest way to avoid it is to insist, at the beginning, that the family of lines be nonempty. There being at least one line, there must exist at least three points: two on the line and one off.]

Our versions of Axioms I-III are deliberately opaque. One of Hilbert's many contributions was to pin down the nature of mathematical abstraction. In an abstract geometry, a point is not an ink dot on a paper or a very small sphere. A point is anything that belongs to a set equipped with what the axioms require. In a similar way, we said earlier that a group can be made up of numbers, permutations, transformations—the kind of objects is immaterial, as long as the ensemble (set, members, operation) conforms to the group axioms. By agreeing to accept the abstraction, we lose the "hidden assumptions" that are inseparable from our mental pictures of points, lines, and other geometric objects. For a particle of evidence, observe that the axioms do not mention the word "straight."

There are not too many theorems you can deduce from those three axioms, but we can begin to see that things that qualify as points and lines behave a little as we would like.

Theorem 1. If two lines intersect, then they have exactly one point in common.

We are more than happy to take advantage of the two meanings of "intersect": the set theory sense of having a common point, and the geometric sense of crossing.

> Say different lines $\mathscr{L}$ and $\mathscr{M}$ have a common point P. Let Q be any other point of $\mathscr{L}$. (Why must such a Q exist?) By Axiom 1, $\mathscr{L}$ is the only line crossing both P and Q. Since $\mathscr{M}$ crosses P and $\mathscr{M} \neq \mathscr{L}$, $\mathscr{M}$ cannot cross Q. Therefore no point other than P is on both lines.

Theorem 2. If P and Q are two points and R is not on the line they determine, then the three points are **noncollinear** (no line crosses all three).

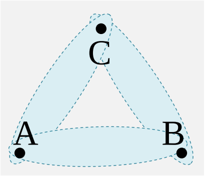
Proof is Exercise 1. Henceforth, we write PQ for the line determined by unequal points P and Q.

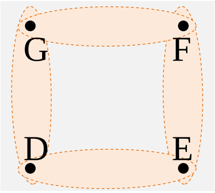
Theorem 3. In the plane, there must exist three noncollinear points and three nonconcurrent lines.

> The family of lines is required to be nonempty. Let $\mathscr{L}$ be some line. By Axiom 2, there must exist two points S and T on $\mathscr{L}$. By Axiom 3, there is at least one point U off $\mathscr{L}$. By Theorem 2, the points S, T, and U cannot be collinear. Those are three noncollinear points.
>
> The three lines ST, SU, and TU cannot be **concurrent**—cannot all have a point in common. They are three different lines; if two were the same, then S, T, and U would be collinear. By Theorem 1, we know S is the only point shared by ST and SU. It cannot be on TU, because then the three would be on that line. Hence no point is on all three lines; we have three nonconcurrent lines.

All this talk about two points and three points sounds like a case of terribly blinkered vision. In fact, a planar geometry might have only three points.

> **Example 1.** Let Σ = {A, B, C} be the plane. To specify a geometry, you have to name the lines. Each line requires at least two points and cannot have all three, and each pair of points has to be on some line. Those conditions leave no choice: The subsets {A, B}, {A, C} and {B, C}have to be the lines. Check that this designation obeys the axioms.
>
> Separately, picture it. At right, the dots represent points and the blue ovals represent lines. Most important is the explicit understanding that only A, B, and C—not any other "points" in the ovals—are points.

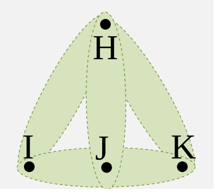




Example 2. Next let Π = {D, E, F, G} and let the orange ovals at left specify the lines. Before you go to the next paragraph, decide why the picture does not specify a planar geometry.

Now add, to the four lines shown, the two subsets {D, F} and {E, G}. Then check that the completed specification obeys the axioms.

> **Example 3.** No doubt you can see that making every pair of elements a line always yields a geometry. But there are alternatives. At right we have four points again, but just four lines (green ovals). Check that the setup satisfies the axioms.



Finally, let us see an example made up of algebraic elements, not geometric. You may, and doubtless should, visualize it in a familiar way.

Example 4. Let the "points" in Σ be the ordered pairs (x, y) of real numbers with

$$x^2 + y^2 < 100.$$

By a "line," we mean those pairs in Σ that satisfy—the **solution set** in Σ of—some linear equation

$ax + by = c,$ a, b, and c fixed real numbers with a and b not both zero,

that has some solutions in Σ. (Geometrically, it is of course Klein's model.) We need to check that the axioms hold. Moreover, we must do it algebraically, not in terms of the Cartesian picture.

Axiom 1: Take distinct points in Σ, like P = (1, 2) and Q = ($\sqrt{5}$, $\sqrt{6}$). The equations

$$A1 + B2 - 1C = 0,$$
$$A\sqrt{5} + B\sqrt{6} - 1C = 0,$$

in the variables A, B, C, constitute a homogeneous linear system with three unknowns and two independent equations. The system necessarily has a one-parameter solution. Here, the solution can be described by

$B = t$ arbitrary, $A = -t\,(2 - \sqrt{6})/(1 - \sqrt{5})$, $C = A + 2B = -t\,(2 - \sqrt{6})/(1 - \sqrt{5}) + 2t$

(Exercise 4). [If the linear-systems language is foreign to you, just eliminate C from the original equations; then keep going under the assurance that this instance is typical.] Therefore P and Q are on at least one line, the solution set of the $t = 1$ equation

$$-1(2 - \sqrt{6})/(1 - \sqrt{5})\,x + 1y = -1(2 - \sqrt{6})/(1 - \sqrt{5}) + 2(1).$$

(If you have no reason to recognize that equation, rearrange it to the point-slope form

$$y - 2 = (2 - \sqrt{6})/(1 - \sqrt{5})\,[x - 1].)$$

They are also on the $t = 2$ line given by

$$-2(2 - \sqrt{6})/(1 - \sqrt{5})\,x + 2y = -2(2 - \sqrt{6})/(1 - \sqrt{5}) + 2(2),$$

among others. But all those are the same line. Their equations are equivalent, except for the equation from $t = 0$, an equation that is not allowed. Therefore they all have the same solution set; they represent the *one* line that crosses P and Q.

Axiom 2: Suppose (u, v) is one point on a line: (u, v) is one solution in Σ of $ax + by = c$. Put some number between $u^2 + v^2$ and 100, say

$$u^2 + v^2 < 99.99.$$

Then

$(U, V) = (u + 10^{-4}b/\sqrt{[a^2 + b^2]},\ v - 10^{-4}a/\sqrt{[a^2 + b^2]})$ (Is that denominator legal?)

is another solution (Verify!), and

$$U^2 + V^2 = u^2 + v^2 + \Big[2u10^{-4}b/\sqrt{[a^2 + b^2]} + 10^{-8}b^2/[a^2 + b^2] - 2v10^{-4}a/\sqrt{[a^2 + b^2]} + 10^{-8}a^2/[a^2 + b^2]\Big].$$

Because u and v are between -10 and 10, the $\big[$bracketed$\big]$ quantity has absolute value no more than

$$[20(10^{-4})\,1 + 10^{-8}\,(1) + 20\,(10^{-4})\,1 + 10^{-8}\,(1)] < 10^{-2}.$$

Hence $U^2 + V^2 < 100$. The equation has a second solution in Σ, and the line has a second point.

Axiom 3: If $ax + by = c$ is the equation of a line, then it has a solution (r, s) with say

$$r^2 + s^2 < 99.99.$$

If $a \neq 0$, then $(r + 10^{-4}a/\sqrt{[a^2 + b^2]}, s)$ is still in Σ (as in the previous paragraph) and does not solve the equation. (Check!) The same is true for $(r, s + 10^{-4}b/\sqrt{[a^2 + b^2]})$ if $b \neq 0$. The equation's solution set does not fill Σ; no line has all the points.

Exercises IX.D.4a: In Exercises 1-3, prove that in a planar incidence geometry:

1. If R is not on the line PQ, then no line crosses all three points.
2. Given two lines, you can find a third line that intersects both.
3. If A and B are unequal points and C is not on line AB, then A is not on line BC.
4. Characterize the simultaneous real solutions of the system

$$\begin{array}{llll} A1 & +B2 & -1C & = 0, \\ A\sqrt{5} & +B\sqrt{6} & -1C & = 0. \end{array}$$

b) multiplanar incidence

Axioms 1-3 specify how lines relate to points. There are five more **incidence axioms**, those that cover what **Merzbach** (page 558) calls "being on" and "being in." Three of them rule how planes—we will end up with multiple planes—relate to points. Another covers how planes relate to lines, and the eighth covers how planes relate to planes.

We move now to **a**(**n incidence**) **geometry**, a set equipped with *two* families of subsets. The members of the set are **points**. In one family, required to be nonempty, we have **lines**, which must obey Axioms 1-3. In the other family, we have **planes**. The package must satisfy Axioms 4-8.

Axiom 4. Any three noncollinear points are on one and only one plane. (Three noncollinear points determine a plane.)

Extending the geometric usage, we say a point belonging to a plane **is on** the plane, and the plane **crosses** the point. A line whose points are in a plane **lies in** the plane, and the plane **contains** the line.

Just from Axioms 1-4, we can see that every point must be on at least two lines and one plane.

> Points must exist: Somewhere, there is a line $\mathcal{L}$, required by Axiom 2 to cross points Q ≠ R.
>
> Let P be any point. It could be that P is on $\mathcal{L}$. Then Axiom 3 promises a point T off $\mathcal{L}$, which means that PT ≠ $\mathcal{L}$ is a second line crossing P. Further, one of Q or R is different from P, say P ≠ Q. By Theorem 2, P, Q, and T are not collinear. By Axiom 4, some plane crosses all three.
>
> Alternatively, P may be off $\mathcal{L}$. In that case, P, Q, and R are noncollinear (Theorem 2). Some plane has to cross them (Axiom 4), and PQ and PR have to be different lines crossing P (Reason?).

Axiom 5. In every plane, there exist three noncollinear points.

Axiom 6. If a plane crosses as many as two points on one line, then it contains the line.

Theorem 4. Any of the following combinations determines a plane:
a) a line and a point off the line;
b) two intersecting lines;
c) two parallel lines (lines lying in the same plane and not intersecting).

> a) Suppose P is off $\mathcal{L}$. The latter must have unequal points Q and R. By Theorem 2, P, Q, and R are not collinear. By Axiom 4, some unique plane Π crosses them. By Axiom 6, since Π crosses Q and R from $\mathcal{L}$, Π must contain $\mathcal{L}$. No other plane can cross P and contain $\mathcal{L}$, because any such plane would cross P, Q, and R. Thus, one and only one plane crosses P and contains $\mathcal{L}$.
>
> b) and c): Exercises 3 and 4.

The situation is getting more familiar. Every line has to lie in some plane, and cannot be any whole plane. The second part is just Axiom 5. For the first part, let $\mathcal{L}$ be a line. Some point T has to be off $\mathcal{L}$, and Theorem 4a says that some plane must cross T and contain $\mathcal{L}$. The hierarchy is what we expect: Points must be on lines, which must lie in planes but not fill them.

Axiom 7. There exist four **noncoplanar** points (four points no single plane crosses).

Even without our convention that “four” means “four distinct”, these four do have to be unequal. Call them A, B, C, D. If two were equal, say A = B, then A = B, C, and D would either be in one line, which would lie in some plane; or would be noncollinear, in which case Axiom 4 would put them on some plane. Moreover, by the same reasoning no three of them can be collinear.

Theorem 5. There must exist at least four points, six lines, and four planes. (Exercise 5)

The numbers in Theorem 5 are the biggest we can guarantee at this time. Picture a tetrahedron, and make a geometry with each vertex as a point, each pair of vertices (and *not* the connecting edge) as a line, and each threesome of vertices (and not their face) as a plane. Then Axioms 1-7 (and Axiom 8 below) are satisfied, and the numbers in the theorem are exact. The same statement holds for the numbers in the next theorem, with the same example as evidence.

Theorem 6. Every point is on at least three lines, and every line lies in at least two planes.

> Pick any point P. We use the points A, B, C, D named above.
>
> It could be that P is on the line AB. In that case, neither C nor D can be on the same line, because no three of A-D are collinear. Hence PC is a second line through P. That line cannot cross D: If it did, then by Theorem 4b those lines, intersecting at P, would lie in a plane crossing A, B, C, D. That makes PD a third line crossing P.
>
> If instead P is off line AB, then PA and PB are unequal lines. (Why?) By Theorem 4b, some plane contains them. That plane cannot cross both C and D: If it did, then the five points would be coplanar. Whichever is off the plane determines a third line through P.
>
> The proof for the line is similar (Exercise 6).

Axiom 8. If two planes have any point in common, then they have at least two points in common.

Theorem 7. If two planes intersect, then their intersection is a line.

There you see that sometimes the language of sets pays off. You also see how much incidence geometry, which we developed abstractly, without pictures, conforms to our mental picture of points, lines, and planes. For Theorem 7, for example, we often compare planes to the surfaces in a typical rectangular room. Such planes may fail to meet, like the floor and ceiling; but if they meet, then they meet along a the line, the way a wall and ceiling meet along the top of the wall.

> Suppose planes $\Pi \neq \Sigma$ have a point P in common. By Axiom 8, they must also cross $Q \neq P$. Each of Π and Σ crosses two points of line PQ. By Axiom 6, each plane contains PQ. Consequently the intersection $\Pi \cap \Sigma$ contains PQ.
>
> [It is essential that you understand that the proof is not done. We now know that PQ *is a subset* of the intersection. We have to show that PQ is *all* of the intersection.]
>
> Suppose R is a point in $\Pi \cap \Sigma$. Then R cannot be off PQ: If it were, then by Theorem 4a only a single plane could contain PQ and cross R; both Π and Σ contain PQ and cross R. Hence R has to be on PQ. We have shown that PQ fills the entire intersection.

Exercises IX.D.4b: In an incidence geometry, prove:

1. If a point is on a plane, then the plane contains a line that does not cross the point.
2. In every plane, it is possible to find three nonconcurrent lines.
3. Two intersecting lines determine a plane.
4. If two lines lie in one plane and do not intersect, then that plane is the only one that contains both lines.

5. There must exist at least four points, six lines, and four planes. [See the information right above Theorem 5.]
6. Any line lies on at least two planes.

c) order

The order axioms start to fill in lines and planes. An **order geometry** is an incidence geometry in which there is defined a ternary relation called **betweenness**.

> We have met binary relations. A **binary relation** in a set picks out ordered pairs of members. In the set of natural numbers, we defined the relation < of "being smaller than" (section IX.C.1c). That relation picks out the ordered pairs (3, 7) and (3, 5), because
>
> $3 < 7$ and $3 < 5$,
>
> but not (5, 3) nor (6, 6). In a similar way, a **ternary relation** in a geometry picks out ordered *triples* (P, Q, R) of points; for such triples, our betweenness relation says that Q **is between** P **and** R.

The betweenness relation must satisfy four axioms.

Axiom 9. It applies only for points in a line: If Q is between P and R, then P, Q, and R are collinear.

We will need to say "A is between B and C" so frequently that we will abbreviate it by $\underline{BAC}$.

Axiom 10. It has left-right symmetry: If $\underline{PQR}$, then also $\underline{RQP}$.

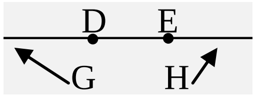
Axiom 11. It obeys the trichotomy (compare section IX.B.4a): If A, B, and C are points on one line, then *exactly one* of the following is true: $\underline{ABC}$, or instead $\underline{ACB}$, or instead $\underline{BAC}$.

Notice that A itself is never between A and B; that would violate the trichotomy.

Axiom 12. If D and E are two points, then there must exist a point F between them.

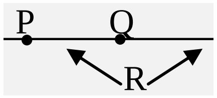
It is immediate from Axiom 12 that every line is an infinite set (Exercise 1). There are two other inferences that deal entirely with betweenness, but require later axioms to prove. We will write them in the next proposition because they are useful. We will prove part (a) later. Part (b) is hard to prove—Hilbert thought it had to be an axiom (refer to Wikipedia®)—and we will not try.

Proposition 1. a) If D and E are two points, then there exist points G and H such that $\underline{GDE}$ and $\underline{DEH}$.
b) Given four points on a line, it is possible to label them P, Q, R, and S so that they appear **in that order**, meaning $\underline{PQR}$, $\underline{PQS}$, $\underline{PRS}$, and $\underline{QRS}$.



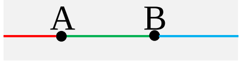
In Proposition 1, part (a) calls for points placed as shown at right. Relative to D and E, we will say that H **is beyond** E and G is beyond D. Part (b) does with betweenness something like what the generalized associative law does with operations. It says that any four points on a line have to be in some order. It can be extended to larger numbers of points; view Exercise 3. We will signify its four betweenness statements by writing $\underline{PQRS}$.

(i) separation on a line



Take two points P and Q. The points between P and Q—which by Axiom 9 are on line PQ—together with P and Q, constitute a **line segment**. We will denote it by [PQ] (a nonstandard symbol mimicking closed interval notation). On the line PQ, we will say Q and the third point R **are on the same side of** P if [QR] does not cross P—in other words, if P is not between Q and R. By Axiom 11 one of the three has to be between the other two; we see that Q and R are on the same side of P iff $\underline{PRQ}$ or $\underline{PQR}$, as the figure suggests. [Having warned against the dangers of drawing inferences from pictures, I offer this advice for the rest of the chapter: *Always draw a picture*.]

Theorem 8. A point on a line separates the line into exactly two **sides**. In symbols: If A is on ℒ, then the remaining points of ℒ fill up two nonempty disjoint subsets *L* and *R*, such that any two points in *L* are on the same side of A, any two points from *R* are on the same side of A, and no point from *L* is on the same side as any point from *R*.



> Assume A is on ℒ. By Axiom 2, there is B ≠ A on ℒ. Let *R* consist of the points of ℒ on the same side of A as B, plus B itself. Let *L* take up the points not on the same side of A as B. By the discussion above, *R* has the points between A and B (green in the figure at right), B, and the points beyond B (blue). That leaves *L* with the points beyond A (red). Those sets are nonempty by Proposition 1a, are disjoint by the trichotomy, and account for all the points of ℒ other than A.
>
> If P ≠ Q are in *L*, then they are also different from A and B. By Proposition 1b, those four points have to be in some order. We know $\underline{\text{PAB}}$ is part of the order, because P is beyond A. Neither $\underline{\text{PAQB}}$ nor $\underline{\text{PABQ}}$ is allowed, because Q is also beyond A. The four points must line up as either $\underline{\text{PQAB}}$ or $\underline{\text{QPAB}}$. In either case, A is not between P and Q; P and Q are on the same side of A.
>
> Assume now S ≠ T are in *R*. It might be that one of them is B, say S = B. Then by the definition of *R*, T is either between A and S = B or beyond S. Either way, $\underline{\text{SAT}}$ is prohibited by the trichotomy; S and T are on the same side of A. It might instead be that S and T are both different from B. Then A, B, S, T are four points. The first three are arranged $\underline{\text{ASB}}$ or $\underline{\text{ABS}}$, by the composition of *R*. In neither of those can T precede A, for the same reason. Therefore the order is $\underline{\text{ATSB}}$, $\underline{\text{ASTB}}$, or $\underline{\text{ASBT}}$. All of those have A outside of [ST]. Again S and T are on the same side of A.
>
> Finally, suppose U is in *L* and V is in *R*. We know $\underline{\text{UAB}}$. One possibility is V = B, so that $\underline{\text{UAV}}$. The others are $\underline{\text{AVB}}$ and $\underline{\text{ABV}}$, which force $\underline{\text{UAVB}}$ or $\underline{\text{UABV}}$. In both of those, U and V are on what we may now call **opposite sides of** A.

Exercises IX.D.4c(i): In an order geometry, prove:

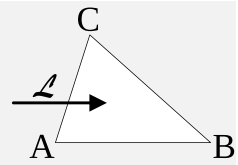
1. Every line—indeed, every line segment—crosses an infinity of points.
2. The intersection of two segments is empty, or has a single point, or is a segment.
3. Any five points on one line have to be in some order. (Clearly the argument suggests the induction proof for six or more points.)

(ii) separation in a plane

Now we add an axiom and show that a line will separate a plane.

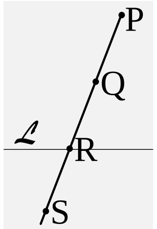
Let A, B, and C be noncollinear. They determine a plane Π. (Reason?) The three segments [AB], [AC], [BC] are subsets of Π. (Why?) Their union is a **triangle**, of which the segments are the **sides** and the points are the **vertices**.

Axiom 13. (Pasch’s Axiom) If a line in the plane of a triangle crosses one side but not the vertices, then it must cross just one of the other sides. In symbols: Assume ℒ lies in the plane of noncollinear points A, B, and C, and does not cross any of them; if ℒ intersects [AC], then it must intersect [AB] or [BC], *and not both*.



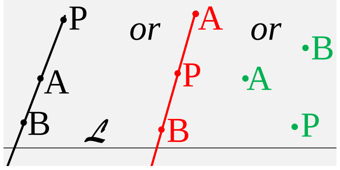
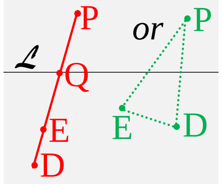
In the corresponding picture, ℒ enters the triangle by crossing a point between A and C. The axiom says that it must exit by crossing a point on precisely one of the other sides of the triangle (and necessarily not at B).

[Moritz Pasch (1843-1930) was a pioneer in axiomatization in general, and especially of geometry. Axiom 13 postulates one of those hidden assumptions we have decried. Pasch insisted that you have to draw conclusions—for example, in Axioms 11 and 13—entirely from axioms (and consequent theorems), and not from consideration of physical situations as in the last figure. (Consider that in the spherical Riemannian geometry, section IX.D.1b(ii), trichotomy fails: Each of three points on a line is between the other two.) In this insistence, he was forerunner to Peano (see "formal system" in section IX.C.1b) as well as Hilbert. Consult O'Connor and Robertson at St Andrews.]



Take now a line lying in a plane. Points A ≠ B in the plane but off the line **are on the same side of the line** if [AB] does not intersect the line. In the figure at right, P is off line ℒ. Choose a point R on the line. By Axiom 12, there must exist Q between P and R. From $\underline{\text{PQR}}$, we deduce that R is not on [PQ]. No other point of ℒ can be on [PQ], because R is the only point of ℒ on line PQ. Therefore P and Q are on the same side of ℒ. By Proposition 1a, there exists S beyond R on PQ. Since [PS] crosses R, P and S are not on the same side of ℒ. Those two sides are all there is.

Theorem 9. A line lying in a plane separates the plane into exactly two **sides**. In symbols: If ℒ lies in Π, then the rest of the points of Π break into nonempty disjoint subsets *S* and *T* with points of *S* on the same side of ℒ, points of *T* on the same side of ℒ, and no point from *S* on the same side of ℒ as any point in *T*.

> Let ℒ lie in Π. There has to be a point P off the line in Π. Let *S* comprise the points in Π on the same side of ℒ as P, *T* the points not on the same side. We saw above that *S* and *T* are nonempty, and by definition they are disjoint and take in all the points off ℒ.
>
> Suppose A and B are two points in *S*. First, it could be that one of them is P, say A = P. In that case, [AB] = [PB] cannot cross ℒ, because B is on the same side as P. Then A and B are on the same side. Second, it could be that A, B, and P are unequal collinear points. In that case, by trichotomy one of them is between the other two. If it is $\underline{\text{PAB}}$ (near left in the figure), then [AB] is a subset of [PB] (Exercise 1a). Since the latter has no points of ℒ, [AB] has none; A and B are on the same side. It goes similarly with $\underline{\text{PBA}}$. If instead it is $\underline{\text{APB}}$ (center), then
>
> [AB] = [AP] union [PB]   (Exercise 1b).
>
> Neither of the latter two intersects ℒ. That means [AB] does not intersect ℒ, and A and B are on the same side of the line. Third, it could be A, B, and P are not collinear (right). Line ℒ does not cross any of them, and intersects neither segment [AP] nor [BP]. By (the contrapositive of) Pasch's axiom, it cannot intersect [AB]. Therefore A and B are on the same side of ℒ.
>
> 
>
>
> Look next at two points D and E in *T*. They are certainly unequal to P. First, it could be that D, E, and P are collinear (left half in the figure at right). Their order cannot be $\underline{\text{DPE}}$. If that were true, then ℒ would intersect line DE at one point between D and P and at a second point between E and P. We must have $\underline{\text{PED}}$ or $\underline{\text{PDE}}$. We treat either the same way; work with $\underline{\text{PED}}$. There must be a point Q on ℒ with $\underline{\text{PQE}}$. The only place to fit Q into $\underline{\text{PED}}$—as Proposition 1b requires—while maintaining $\underline{\text{PQE}}$ is $\underline{\text{PQED}}$. With that order, [ED] has no point of ℒ, because ED and ℒ share only Q. That tells us E and D are on the same side. Second, it might be that P, D, and E are noncollinear. In that case, PDE is a triangle. The line ℒ does not cross P, D, or E, and intersects sides [PE] and [PD]. By the "not both" provision in Axiom 13, ℒ does not cross [DE]. Therefore D and E are on the same side of the line.
>
> 
>
>
> That leaves the case of A in *S* and D in *T* for Exercise 2.

It is worthwhile to see by example that Pasch's axiom is independent of the previous ones and essential for Theorem 9. We will use a modification of three-dimensional Cartesian space.

Example 5. Look at the set **S** of nonzero ordered triples $(r, s, t) \neq (0, 0, 0)$ of *rational* numbers. Define a **plane** as the solution set in **S** of a linear equation

$ax + by + cz = d$, a, b, c, and d rational, with a, b, c not all zero.

You can check that the process of solving two such equations simultaneously yields one of three results: The equations may be equivalent, representing the same plane; they may be inconsistent, so that their planes do not intersect; or they may have an infinity of simultaneous solutions, given by one parameter. In the last case, if one solution is (u, v, w), then all the solutions are given by

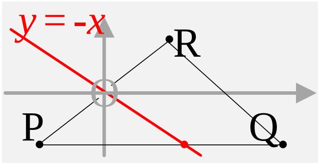
$x = u + At$, $y = v + Bt$, $z = w + Ct$, t arbitrary rational,

in which A, B, and C are three rationals not all zero. Define a **line** as the set of nonzero triples given by such a form. We can then verify that Axioms 1-8 hold.

On the line given by that form, we say of three nonzero points that

$(u + At, v + Bt, w + Ct)$ **is between** $(u + Ar, v + Br, w + Cr)$ **and** $(u + As, v + Bs, w + Cs)$

if either $r < t < s$ or $s < t < r$. From that definition, we can check that Axioms 9-12 are satisfied. It is also easy to check that Proposition 1 holds. What fails is Pasch's axiom.



In the xy-plane, given by

$0x + 0y + 1z = 0$,

look at the points P(-1, -1, 0), R(1, 1, 0), Q(3, -1, 0). In that plane, PQ is given by $y = -1$. (Officially it is

$x = -1 + 1t$, $y = -1 + 0t$, $z = 0 + 0t$;

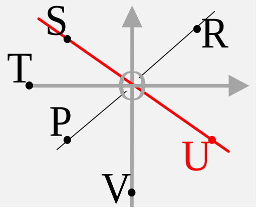
let us agree, in the figure and what follows, to suppress the z-coordinate.) PR is given by $y = x$, QR by $y = -x + 2$. The (red) line with $y = -x$ intersects [PQ] at (1, -1) (red dot). But it intersects neither [PR] nor [QR]. The system

$y = -x$ and $y = -x + 2$

has no simultaneous solution, and the only simultaneous solution to

$y = -x$ and $y = x$

is not a member of **S**. That violates Axiom 13.

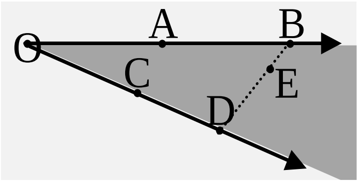


As a result, PR does not separate the plane. It is clear that S(-1, 1) and T(-2, 0) are on the same side of PR. So are U(1, -1) and V(0, -2). Surprisingly, S and U are on the same side of PR; we just noted that line SU does not intersect line PR. So we have T on the same side as S, S on the same side as U, and U on the same side as V. But T and V are on opposite sides: TV is given by $y = -x - 2$, and therefore segment [TV] intersects PR at P.

If you allow (0, 0, 0)—if you take *all* the triples in $\mathbf{Q}^3$ instead of just those in **S**—then the geometry will satisfy Pasch's axiom, even though any line will still have a hole at every real triple with an irrational coordinate. In that case, the simultaneous solutions of the equations of our lines, being rational triples, will always be points in the geometry.

(iii) rays and angles

Take point O on some line. We define a **ray** or **half-line** as the set consisting of O and the points of the line on one side of O. If A is on that side, we use [OA⟩ to denote the ray. Notice that if B is another point on the ray, as in the figure below right, then [OA⟩ = [OB⟩.



Now assume C is not on line OA. The union of the two sets [OA⟩ and [OC⟩ is an **angle**. We denote the angle by ∠AOC, and call the rays **sides** of the angle, O the **vertex** of the angle. By the definition,

∠AOC = ∠COA.

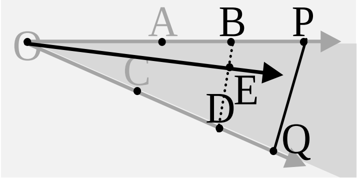
Also, if D ≠ O is on [OC⟩, then

∠AOC = ∠AOD = ∠BOC = ∠BOD. (They are all the same set.)

Keep in mind that ∠AOC has the points on the black half-lines, not those in the gray area.

The gray area has a special name. We have excluded “straight angles”; we required O, A, and C to be noncollinear. Then OA and OC are unequal intersecting lines. By Theorem 4b, a unique plane contains both lines, and therefore contains the angle. In that plane, all the points of [OA⟩ save O are on the same side of line OC: No point of OC can be on segment [AB], because only O is on lines OC and AB. Similarly, all the points of [OC⟩ except O are on the same side of line OA. We define the **interior** of ∠AOC as the set of points E in the plane that are between some points B ≠ O on [OA⟩ and D ≠ O on [OC⟩. The interior is not empty, since it has all the points between A and C. If E is in the interior, then E is on A and B’s side of OC in the plane, and on C and D’s side of OA. The converse is also true, but it is one of those things we need later axioms to prove. The converse is that if P is simultaneously on A’s side of OC and on C’s side of OA, then there are B and D on the respective rays with $\underline{\text{BPD}}$. One thing we can prove now touches on the very first “unstated assumption” we met (section III.A.5b).

Theorem 10. Suppose E is in the interior of ∠AOC. If segment [PQ] **crosses the angle**—P ≠ O is on [OA⟩ and Q ≠ O is on [OC⟩—then ray [OE⟩ has to intersect [PQ].



> By definition, E is between B on [OA⟩ and D on [OC⟩. If [BD] = [PQ], then E is on [PQ]. If B = P and D ≠ Q, then B, D, and Q are noncollinear, because B is not on line OC. Line OE crosses none of those points, because it shares only O with lines OA and OC. We know it intersects side [BD] of triangle BDQ at E. By Pasch’s axiom, OE has to intersect side [BQ] = [PQ]. The intersection, necessarily in the interior of ∠AOC, must be on ray [OE⟩. Finally, if B ≠ P and D ≠ Q, then the same argument shows that OE has to intersect side [BQ] of triangle BDQ, wherefore it has to meet the remaining side [PQ] of triangle BQP.

Exercises IX.D.4c(iii): In an order geometry, prove:

1. a) If $\underline{\text{PAB}}$, then any point between A and B is between P and B.
   b) If $\underline{\text{APB}}$, then [AB] is the union of the subsets [AP] and [PB].
2. The rest of Theorem 9: If A is on the same side of $\mathcal{L}$ as P and D is not, then A is not on the same side of $\mathcal{L}$ as D.
3. Assuming that the interior of ∠AOC is the intersection, in the plane of the angle, of A’s side of line OC and C’s side of OA:
   a) The interior is **convex**: If E and F are in the interior, then [EF] is contained in the interior.
   b) The **exterior** (the set of points in the plane not on the angle or in the interior) is **arc-connected**: If G and H are in the exterior, then there is a **broken line from** G **to** H (specifically, a union of segments [GP], [PQ], [QH]) that is contained in the exterior.
   c) If E is in the interior and G in the exterior, then any broken line (any union of a finite number of segments) from E to G must intersect the angle.

d) congruence

You can see that the proofs are becoming increasingly elaborate. That is always the difficulty in trying to give a complete axiomatic foundation to a deductive system as multifaceted as geometry. To avoid the complication, we will give some statements only partial arguments. Below we list the remaining axioms and sample some especially familiar statements that follow from them.

The last related group of axioms concerns congruence. To turn an order geometry into a **congruence geometry**, we require that there exist two binary relations. One relates line segments, the other relates angles. Both are called **congruence** and use the symbol ≅. They must separately obey four axioms, and a fifth axiom connects them. [Recall (section IX.A.2c(ii)) how in a field, the only axiom that connects addition and multiplication is the distributive law.]

Axiom 14. Each congruence is an equivalence relation (refer to Exercise VIII.C.2a:3). In symbols:

$[AB] \cong [AB]$ for every segment;
if $[AB] \cong [CD]$, then $[CD] \cong [AB]$;
if $[AB] \cong [CD]$ and $[CD] \cong [EF]$, then $[AB] \cong [EF]$;

and similarly for congruence of angles.

(i) segments

Axiom 15. Given A ≠ B and some line crossing point C, there exist on that line a unique D on one side of C, and a unique E on the opposite side, such that

$$[CD] \cong [AB] \cong [CE].$$

That is a familiar idea. It demands that you can reproduce a “length”—we will call it “lay off a segment”—along any line, in either direction from any of its points.

Example 6. Return to the set $\mathbf{Q}^3$ of ordered triples of rational numbers, including (0, 0, 0). Use the plane, line, and betweenness definitions given in Example 5. We noted there that the order geometry so established satisfies Axioms 1-13. In it, we could define congruence of segments by reference to the distance formula. That is, if $A = (a_1, a_2, a_3)$, $B = (b_1, b_2, b_3)$, and so on, we define

$[AB] \cong [CD]$ to mean

$$(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 = (c_1 - d_1)^2 + (c_2 - d_2)^2 + (c_3 - d_3)^2.$$

That definition clearly satisfies Axiom 14.

It fails Axiom 15. Put A at (1, 1, 1) and B at (2, 2, 2). You cannot lay off [AB] anywhere along the x-axis. The axis is given by

$$x = 0 + 1t, \qquad y = 0 + 0t, \qquad z = 0 + 0t, \qquad t \text{ rational}.$$

There are no rational t and s for which

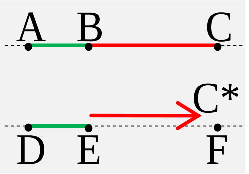
$$(1t - 1s)^2 + (0t - 0s)^2 + (0t - 0s)^2 = (2-1)^2 + (2-1)^2 + (2-1)^2.$$

Axiom 16. Given A, B, and C in that order on a line, and D, E, and F in that order on a (not necessarily different) line, if

$$[AB] \cong [DE] \quad \text{and} \quad [BC] \cong [EF], \qquad \text{then} \qquad [AC] \cong [DF].$$

Clearly the axiom is the “equals added to equals …” statement. Where such **additivity** prevails, subtractivity [not really a mathematical word] also holds. To prove that:

First we prove that “the whole cannot equal a part.” Assume $\underline{ABC}$. Then [AB] cannot be congruent to [AC]. That congruence would violate the uniqueness part of Axiom 15; you would have unequal points B and C on the same side of A reproducing [AB]. See also Exercise 1.



Now assume ABC and DEF, and suppose (illustrated at right)

[AC] ≅ [DF] and [AB] ≅ [DE].

Lay off [BC] from E toward F, ending at C*. We have ABC, DEC*, and

[AB] ≅ [DE] and [BC] ≅ [EC*].

By additivity,

[DF] ≅ [AC] ≅ [DC*]. In view of the previous paragraph, C* cannot be left of F (meaning EC*F) nor right of F (EFC*). Only C* = F is possible, and we conclude

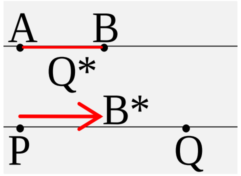
[BC] ≅ [EC*] = [EF].

We can now define relative size of segments. Given segments [AB] and [PQ], lay off a copy of [AB] from P toward Q, ending at B*. There are three places B* could be: between P and Q, at Q, or past Q.

If B* = Q, then the segments [PB*] and [PQ] are one, and

[AB] ≅ [PB*] = [PQ].

If PB*Q (figure below right), then we say [AB] **is shorter than** [PQ]; if PQB*, then [AB] **is longer than** [PQ]. Notice that exactly one of those is true.



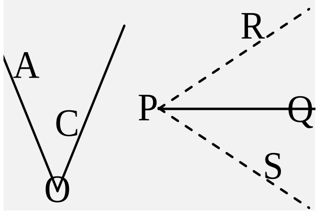
Moreover, [AB] is shorter than [PQ] iff [PQ] is longer than [AB]. In the figure, [AB] is congruent to [PB*] within [PQ]. First, [PQ] cannot be congruent to [AB], because then we would have the whole [PQ] congruent to the part [PB*]. Second, [PQ] cannot be shorter than [AB]. In that case, there would be Q* between A and B with [AQ*] ≅ [PQ]. We could then add a copy of [Q*B] from Q rightward to S (off to the right of the figure). By additivity, we would have

[PS] ≅ [AB] ≅ [PB*].

That would be a whole congruent to a part. We conclude [PQ] is longer than [AB].

The converse is symmetric to our statement.

(ii) angles and triangles



Axiom 17. Given noncollinear points A, O, C, as well as P ≠ Q in some (maybe the same) plane Π, there exist in Π precisely two rays [PR⟩ and [PS⟩, one toward each side in Π of line PQ, such that

∠QPR ≅ ∠AOC ≅ ∠QPS.

Again in words: You can make a copy of any given angle to lie in any specified plane and to either side of a ray that is to be one side of the copy.

Now we connect the two types of congruence.

Axiom 18. Given two (possibly equal) triangles, in which two sides and the **included angle** (where the sides meet) of one are congruent to two corresponding sides and the included angle of the other, the remaining sides must be congruent, and the remaining pairs of corresponding angles must be congruent.

This is, of course, the SAS principle. In symbols, it assumes that ABC and DEF are triangles with

[AB] ≅ [DE],   ∠BAC ≅ ∠EDF,   and   [AC] ≅ [DF].

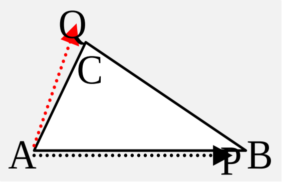
Given that, it requires that the remaining congruences,

[BC] ≅ [EF],   ∠ABC ≅ ∠DEF,   and   ∠ACB ≅ ∠DFE,

follow. When all those congruences hold, we say **the two triangles are congruent**.

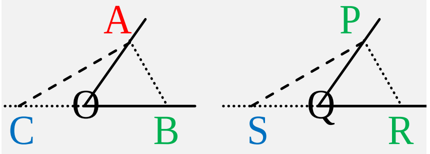
We claimed long ago (section III.A.1) that SSS was the more fundamental principle about equal size and shape. Clearly, though, we need *something* to connect segment and angle congruence; SSS would not do that. Let us see some interesting consequences of SAS; SSS is among them.

Theorem 11. a) The base angles in an isosceles triangle are congruent (Exercise 2).
b) (ASA) If two angles and the included side of one triangle are congruent to corresponding angles and the included side of another, then the triangles are congruent.

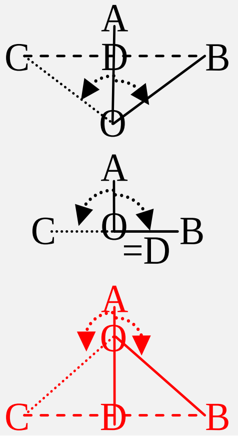
> b) Assume that in triangles ABC and DEF, we know
>
> ∠CAB ≅ ∠FDE, [AB] ≅ [DE], and ∠ABC ≅ ∠DEF.
>
> Lay off [DE] from A toward B, reaching P. Point P cannot land before B or beyond B; if it did, we would have the whole congruent to a part. Thus, P coincides with B. At A, draw a ray (red at right) that duplicates ∠EDF toward the C side of the plane. That ray coincides with [AC⟩; otherwise it and [AC⟩ would give two duplicates of ∠EDF, violating the uniqueness part of Axiom 17. Take Q along the ray to make [AQ] ≅ [DF]. From
>
> 
>
>
> [AQ] ≅ [DF] ≅ [AC],
>
> we conclude Q = C. Triangles QAP and FDE satisfy SAS by design. But triangle QAP *is* triangle CAB. We have shown that triangles CAB and FDE are congruent.

From ASA, we immediately get the converse of the base-angles theorem (Exercise 3).

Theorem 12. Congruent angles have congruent supplements.

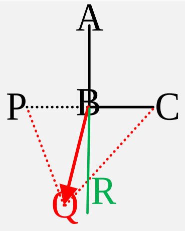
> Given congruent angles (solid black in the figure) with vertices at O and Q, start with A on one of their rays. Lay off [OA] along the other three rays, to make
>
> 
>
>
> [OA] ≅ [OB] ≅ [QP] ≅ [QR].
>
> By SAS, triangles AOB and PQR are congruent. Therefore
>
> ∠OBA ≅ ∠QRP and [AB] ≅ [PR].
>
> Now pick C beyond O along OB, and S beyond Q along QR, such that [OC] ≅ [QS]. By additivity, [BC] ≅ [RS]. By SAS, triangle ABC is congruent to triangle PRS. Therefore
>
> ∠ACB ≅ ∠PSR and [AC] ≅ [PS].
>
> That means triangles ACO and PSQ satisfy SAS. Hence ∠AOC and ∠PQS, supplements to the original congruent angles ∠AOB and ∠PQR respectively, are congruent.

From congruent supplements, it is immediate that vertical angles are congruent (Exercise 4). We can also prove that an angle might be congruent to its supplement. In that case, we call it a **right angle**.

> Take any angle ∠AOB. There is a unique ray (dotted in each third of the figure at right), on the side of line AO opposite B, to make an angle congruent to ∠AOB. Put C on that ray to make [OC] ≅ [OB]. Because B and C are on opposite sides of AO, the segment [BC] must intersect AO at a point D.
>
> 
>
>
> If D is on A's side of O along OA (top third), then triangles ODB and ODC are congruent by SAS, and ∠ODB and ∠ODC are congruent supplements. If instead D = O (middle third), then B, O, and C are on the same line, and ∠AOB and ∠AOC are supplements as well as congruent. The remaining possibility puts D beyond O. In that case (all red in the lowest third) ∠DOB and ∠DOC must be congruent, because they are supplements to the original congruent pair. Therefore triangles DOB and DOC are congruent, and ∠ODB and ∠ODC are congruent supplements.

That argument shows that right angles exist. Are they all the same size?

Theorem 13. Any two right angles are congruent.



Suppose ∠ABC and ∠DEF are right angles. In the figure (which does not show the latter), put P on CB beyond B to make [BP] ≅ [BC]. From ray [BP⟩ toward the side of the plane opposite A, draw the ray [BQ⟩ (red) that makes ∠PBQ ≅ ∠DEF. Last, draw the extension (green) of AB beyond B.

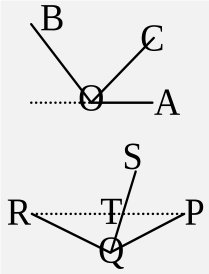
Angle ∠PBQ is by design congruent to ∠DEF, which by definition is congruent to its supplement, which by congruent supplements is congruent to ∠CBQ. (That shows that a copy of a right angle is a right angle. Do the related Exercise 5.) Hence triangles PBQ and CBQ satisfy SAS. That implies ∠BCQ ≅ ∠BPQ.

Line AB enters triangle CPQ and misses P and C. It must therefore intersect one of the other two sides: Either it crosses Q, or else it falls under Pasch's axiom. The intersection has to be on the Q side. Say it is at point R on [QC]. (Similar reasoning applies if R is on [PQ].) Angles ∠CBR and ∠PBR are right angles (Exercise 6). Therefore triangles CBR and PBR fit SAS. We now have

$$\angle BPR \cong \angle BCR = \angle BCQ \cong \angle BPQ.$$

The ray [PQ⟩ must coincide with [PR⟩, and so it intersects CQ = CR at Q = R. That means ∠PBQ is vertical to ∠ABC, whence

$$\angle ABC \cong \angle PBQ \cong \angle DEF.$$



Theorem 14. Assume rays [OC⟩ and [QS⟩ are interior to angles ∠AOB and ∠PQR (as in the figure at right). Assume further that

$$\angle AOC \cong \angle PQS \qquad \text{and} \qquad \angle COB \cong \angle SQR.$$

Then $\angle AOB \cong \angle PQR$.

Clearly the theorem states the additivity of angles. As usual, additivity implies subtractivity. (How would you state the latter?)

To start the proof, put P* along [OA⟩ (figure below right) so that

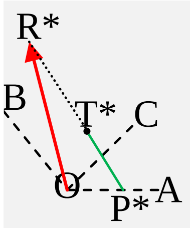
$$[OP^*] \cong [QP].$$

Then choose R* (along the red ray) so that

$$\angle P^*OR^* \cong \angle PQR \quad \text{and} \quad [OR^*] \cong [QR].$$

With those matches, triangles P*OR* and PQR are congruent. Therefore

$$\angle OP^*R^* \cong \angle QPR \quad \text{and} \quad [P^*R^*] \cong [PR].$$



The ray [QS⟩ has to intersect [PR] (Reason?) at a point T (previous figure). Pick T* along P*R* such that [P*T*] (green at right) is congruent to [PT]. Then triangles OP*T* and QPT satisfy SAS. Hence

$$\angle P^*OT^* \cong \angle PQT.$$

By hypothesis, ∠PQT = ∠PQS is congruent to ∠AOC. That means ∠P*OT* and ∠AOC both duplicate ∠PQT. It must be that ray [OT*⟩ coincides with [OC⟩.

Examine next triangles OT*R* and QTR. By the congruence of the big triangles, we have

$$\angle OR^*T^* \cong \angle QRT.$$

By subtraction, we have

$$[T^*R^*] \cong [TR].$$

By congruent supplements, we have

$$\angle OT^*R^* \cong \angle QTR.$$

The triangles are congruent by ASA. Hence

$$\angle T^*OR^* \cong \angle TQR.$$

By hypothesis, the last is congruent to ∠COB = ∠T*OB. Having now

∠T*OR* ≅ ∠T*OB,

we conclude that [OR*⟩ coincides with [OB⟩. Additivity follows:

∠AOB = ∠P*OR* ≅ ∠PQR.

Observe that our definition of angle did not allow what we may not yet call “straight angles.” The statement of Theorem 14 and the argument above depend on the overlying angles’ having an interior. However, check (Exercise 9) that if B is beyond O on line AO and

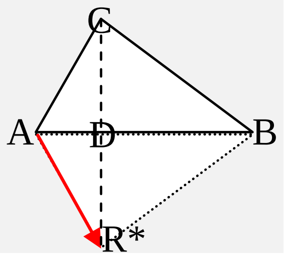
∠AOC ≅ ∠PQS and ∠COB ≅ ∠SQR,

then R has to be beyond Q on PQ. That means we keep additivity if we extend the definition and the congruence: If $\underline{\text{AOB}}$, we say ∠AOB (which as a set is simply AB) is a **straight angle**; and a straight angle **is congruent to** other straight angles and only to those. (Take additivity further in Exercise 10.)

As before, additivity allows us [you] to define relative sizes of angles (part of Exercise 12).

Theorem 15. (SSS) If the sides of one triangle are congruent to corresponding sides of a second, then the triangles are congruent.

Suppose triangles ABC and PQR have SSS. Clearly, we need only match one angle in ABC to a corresponding one in PQR.



Toward the side of line AB opposite C, pick R* so that

∠BAR* ≅ ∠QPR and [AR*] ≅ [PR].

Triangles BAR* and QPR are congruent by SAS. Therefore BAR* and BAC have SSS.

Because C and R* are on opposite sides of AB, [CR*] has to intersect AB at a point D. We deal with the case $\underline{\text{ADB}}$; you get the cases (Exercise 11) where either D is one of A and B, or instead D is beyond one of A and B.

With D between A and B, triangles ACR* and BCR* are isosceles. By base angles,

∠ACR* ≅ ∠AR*C and ∠R*CB ≅ ∠CR*B.

By additivity,

∠ACB ≅ ∠AR*B.

Since the last is congruent to ∠PRQ, we have the angle match we need.

Exercises IX.D.4d: In a congruence geometry, prove:

1. If unequal points B and D are between A and C, then [BD] cannot be congruent to [AC].
2. In an isosceles triangle, the base angles are congruent. (It is up to you to define **isosceles triangle** and **base angles**.)
3. If two angles in a triangle are congruent, then their opposing sides are congruent.
4. Vertical angles are congruent. (Define **vertical angles**.)
5. At a given point on a line lying in some plane, there exists a unique perpendicular. (You need to define **perpendicular**.)
6. If two lines intersect to form a right angle, then all four angles at the intersection (excluding the straight angles) are right angles.
7. Assuming the base of an isosceles triangle has a midpoint, the median to it must be perpendicular to the base. (You need **midpoint** and **median**.)
8. If two lines in a plane are cut by a transversal so as to form congruent a/i angles, then the lines must be parallel. (Define **transversal** and **a/i angles**.)

9. If AOB, and C off line AB satisfies
   $\angle AOC \cong \angle PQS$ and $\angle COB \cong \angle SQR$,
   then $\angle PQR$ is also a straight angle.
10. Additivity still works when the angles to be added sum to more than a straight angle. (Define "sum to more ….")
11. In the last figure, triangles ABC and ABR* (having SSS) are congruent if:
    a) The crossing D is at B; or instead,
    b) It is beyond B.
12. In a triangle, the shorter side is opposite the smaller angle. (Define **smaller angle**.)

e) Euclidean space

Of the remaining three axioms, two are more analytical than geometric. They say that lines are put together like the number line. Remarkably, they do so without any mention of numbers. They are part of what makes it possible to introduce numbers, in particular to define "distance."

Axiom 19. Given two points on a line and some line segment, it must be possible to lay off the segment multiple times along the line, starting at one of the points, so as to get beyond the other. In symbols: Let A ≠ B and C ≠ D; there must exist points $A = A_0, A_1, \ldots, A_n$, in that order toward B's side of A along line AB, such that

$$[A_0A_1] \cong [A_1A_2] \cong \ldots \cong [A_{n-1}A_n] \cong [CD]$$

and A_n is beyond B.

The axiom says that all the points of a line are within reach. That is, enough steps of the size of [CD], no matter how small [CD] is, will get you from any starting A past any B on the given line. Evidently it is the axiom of Archimedes. Notice that it establishes the *existence* of points beyond B; it verifies Proposition 1a.

Axiom 20. Every line must be **complete**. That is, it is impossible to add a point to a line and then *extend* to the enlarged line the incidence, order, and congruence relations from the original line.

We have met "extend" before. In section IX.B.4c(i), we said that Brahmagupta's zero and signed numbers enlarge the set of natural numbers to the set of integers. The operations of integer addition and integer multiplication, on the enlarged set, *extend* the natural-number operations. That is, if m and n are natural, then the results $m + n$ and mn under the *integer* operations are the same as they were under the original natural operations. Axiom 20 says you cannot enlarge a line and make definitions of membership, betweenness, and congruence on the enlarged line that extend the original ones.

[More recently, above Theorem 15 , we enlarged the definition of angle by defining straight angles, and we extended the definition of congruence to those. That example is weaker than Brahmagupta's, in that there is no congruence between the new, straight angles and the original angles.]

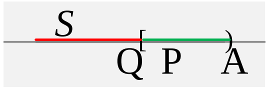
Axiom 20 is invariably called the "completeness axiom." It is a kind of continuum property. We could choose to replace it, as well as Axiom 19, with a single axiom having a continuum form. Look at Dedekind's theorem, section IX.B.4c(ii). We could write **Dedekind's Axiom** as:

Axiom 18.5. Suppose any line $\mathcal{L}$ is partitioned into nonempty disjoint sets L and R, such that no point of L is between two points of R, and vice-versa. Then $\mathcal{L}$ must cross a point C (that we will choose to call the "cut point") with the property that L is one side of C on $\mathcal{L}$ (plus maybe C), and R is the other side of C (plus C iff $\mathcal{L}$ does not have it).

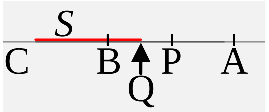
> To show that completeness follows, assume Axiom 20 is false: It is possible to add a new point Z and extend $\mathscr{L}$’s relations. In the enlarged line $\mathscr{L}^* = \mathscr{L} \cup \{Z\}$, take any A ≠ Z. Since the order laws still hold, we can find point B beyond Z. Let *L* consist of the points of $\mathscr{L}^*$ on A’s side of Z, *R* the points on B’s side. Notice that *L* and *R* are subsets of $\mathscr{L}$; neither has Z. By Theorem 8, they are nonempty and disjoint, cover all of $\mathscr{L}$, and have no point of one between two points of the other. They satisfy Dedekind’s hypotheses. But there is no cut point.
>
> That is, there is no cut point in $\mathscr{L}$. (Check that Z is the cut point in $\mathscr{L}^*$.) Points on $\mathscr{L}$ are in either *L* or *R*. If C is in *L*, then there must be P and Q in $\mathscr{L}^*$ with $\underline{\mathrm{PCZ}}$ and $\underline{\mathrm{CQZ}}$. Those points are different from Z; they come from $\mathscr{L}$. That disqualifies C: A cut point cannot have members of *L* on both sides of it. The same argument applies for D in *R*. We have shown Axiom 18.5 is false. By contraposition, Dedekind implies completeness.

More familiarly, we can also establish a kind of least-upper-bound property. Suppose A is a point on line $\mathscr{L}$, and let *S* be a subset of $\mathscr{L}$. We say P ≠ A is a **bound for *S* on the** A**-side** if no point of *S* is on A’s side of P on $\mathscr{L}$.

Theorem 16. Assume A is on $\mathscr{L}$ and the nonempty set *S* (red in the figure at right) has a bound P on the A-side. Then there exists an **extreme bound**, a bound Q on the A-side such that every point between Q and A (green) is also a bound and no point beyond Q is a bound.



> Partition $\mathscr{L}$ into two subsets, *R* holding the bounds on the A-side plus A and the points beyond A, *L* holding the rest. Clearly *R* is not empty. Neither is *L*. If B is in *S*, then the line must have point a point C beyond B (figure at right); C, being left (on P’s side) of A and not a bound, is in *L*. Dedekind’s axiom guarantees a cut point Q.
>
> The cut point cannot be right of P. If it were, then P from *R* and C from *L* would be on the same side of Q; that is not allowed for a cut point. Accordingly, A is an element of *R* right of Q. All the conclusions follow. First, all the points of $\mathscr{L}$ right of Q have to be in *R*. That means all the points between Q and A are bounds on the A-side. Second, Q is a bound. If T is right of Q, then any U between Q and T is a bound, so T cannot belong to *S*. Third, all the points left of Q have to be in *L*, which means they cannot be bounds.



We used the LUB property to prove the axiom of Archimedes in section IX.B.4b(ii). We can do the same here (Exercise 1). For a more elementary result, we can show that any segment can be subdivided into any natural number of equally long parts (Exercise 2).

We can now start talking about lengths and distances. Fix some segment [AB] as the unit of measure. We can make natural lengths by reproducing it. We can make rational lengths by subdividing natural lengths. Where there are rational lengths and Dedekind cuts, there are real-numbers lengths. Finally, we can add signed lengths by measuring to opposite sides along lines.

The axioms are supposed to axiomatize Euclidean geometry. Because Klein’s model, Example 4 in section IX.D.4a, satisfies our Axioms 1-20 but is not Euclidean, we have no choice but to add the parallel postulate. Our last axiom, therefore, is the postulate, in Playfair’s form.

Axiom 21. Let point P be off line $\mathscr{L}$. In the plane they determine [in accordance with Theorem 4a], there must exist exactly one line crossing P and not intersecting (therefore parallel to) $\mathscr{L}$.

Notice that the axiom is entirely about incidence. We could have put it back at (a). It is customary to put it late in the list, maybe to avoid disqualifying Klein too early.

Exercises IX.D.4e

1. Assuming Theorem 16, prove Axiom 19.
2. Assuming Theorem 16, prove that a segment can be divided into *n* congruent pieces. (Hint: Set *n* = 3, define what it means for P to be left of 1/3 of the way from A to B, then proceed.)
3. Assuming Exercise 2, show that any angle has a bisector.
4. Prove our “parallel postulate”: If two lines are parallel, then any transversal forms congruent a/i angles (defined as part of Exercise IX.D.4d:8).

Chapter X. Epilogue

The book ended one page ago. You might wonder why a 2015 history would stop at roughly 1910. The reason is that twentieth-century discoveries in (what became of) geometry and algebra are advanced way beyond a treatment that dreams of staying elementary. You could say the same for number theory, but we will get to an application of it whose importance and elementariness demand a look.

Other than that, we want to take a last look at deductive systems, then see some victories, some defeats, and implications for the future.

Section X.A. Deductive Systems

1. Aristotle and Euclid

The idea of devising a set of axioms from which all knowledge in some area of study follows undoubtedly predates Aristotle (384-322 BCE), who preceded Euclid (dates—even birthplace—unknown, but he was already a famous adult when the first Ptolemy brought him to Alexandria, 323 or later). We have said that Euclid's system became the model for deductive reasoning, never mind our objections that some inferences hang on assumptions not laid out explicitly. Before him, though, Aristotle worked to *define* deductive reasoning.

At the top of an edifice of thought—he had first to write about words and their meanings, then phrases, sentences and their meanings—Aristotle exhibited a set of **rules of inference**. Those are declarations that certain sequences of sentences are (**logically**) **valid** arguments. That is, such **syllogisms** yield true **conclusions** (literally, their last sentences) in every instance where their **premises** (the other sentences) are true, irrespective of what facts (or nonsense) the sentences cover.

> The name **disjunctive syllogism** applies to the form
>
> | | |
> |---|---|
> | *p* or *q* | (Meaning: Sentence *p* is true, or sentence *q* is true, or both are true.) |
> | not *p* | (Sentence *p* is false.) |
> | Therefore *q* | (*It follows* that sentence *q* is true.) |
>
> We have made that kind of inference, as early as when we discussed Egyptian unit fractions (section II.A.2). There, we reasoned:
>
> The biggest-fit method goes on forever, or it terminates.
> The method cannot go on forever (within Exercise II.A.4:6).
> Therefore it terminates.
>
> That termination can happen only when the original fraction has been broken into unit fractions.
>
> Our most productive syllogism has been **universal instantiation**:
>
> Every (member of some set) is {possessed of some property}.
> ⟨Whatever⟩ is (a member of the set).
> Therefore ⟨whatever⟩ is {possessed of the property}.
>
> [In books, the universal (no pun) example of this syllogism is
>
> Every man is mortal.
> Socrates is a man.
> Therefore Socrates is mortal.]
>
> We have reasoned that way, tacitly, time after time. The reason is that almost all our theorems are universal statements. Think of the uncounted times we said triangles were congruent by SAS. [Fifteen; I counted.] Our statement of the SAS axiom had "if-then" form: "Given …, the remaining … must be congruent."). However, the axiom is the universal statement
>
> Every (pair of triangles satisfying SAS) is {a congruent pair}.

> If we establish that
> ⟨Triangles ABC and PQR satisfy SAS⟩,
> then, Aristotle tells us, it *logically follows* that
> {triangles ABC and PQR form a congruent pair}.

The syllogisms in Aristotle's collection are almost all so elementary as to make you wonder why they need stating. Aristotle's work did for them what Pasch's axiom does: It made inference patterns explicitly axiomatic, patterns we would otherwise have to claim to be "self-evident."

2. Added Systems

Newton's laws constitute a set of axioms to make mechanics a deductive system. So do Maxwell's equations for electricity and magnetism. We know now that Newton's mechanics—the laws plus the resulting equations—makes incorrect predictions at molecular scale and in accelerating systems. That merely restricts the domain in which it is *useful*, not its validity. In the other sciences, Dmitri Mendeleev's assumptions and conclusions about the elements (read about the periodic table) make up a wildly successful axiomatic system.

Within mathematics, we saw the nineteenth century's explosion in deductive systems. The works of Bolzano and Cauchy began the axiomatization of analysis. Group theory, flowing from the axioms of Galois, is now just one deductive system within abstract algebra. Riemann's work on surfaces, and Klein's elaboration of it in terms of groups. underlie one deductive system within topology.

The most elementary of the systems is Peano's arithmetic. Consequently we will focus our attention there. (Incidence geometry is elementary also, but it needs twenty axioms, as against Peano's five.)

3. The Limits of Deduction

The ideal system would flow from a set of axioms that is minimal (you cannot spare any of them; they are **independent**), allows us to deduce all facts in its area (it is **complete)**, and never leads to contradictions (it is **consistent**). Our next topic is the stunning discovery that no deductive system worthy of the name can be both complete and consistent: If it is consistent, then there are truths in it that you cannot reach by deduction.

a) formal systems

In section IX.C.1b, in the context of defining addition, we alluded briefly to Peano's attempt to develop arithmetic as a formal system. Building a formal system is remarkably hard. You have to define a vocabulary, specifying which symbols or sequences of symbols can represent objects. Thus, you might allow "1" and "*F*(1)", and exclude ")*F*1=$". You need to define **predicates**, things you can say about the named objects. For example, "is not equal to 1" is something you could allow to be said of *F*(1). You must specify which sequences of symbols constitute **sentences**, so that say

"Every element has a follower."

is allowed (is **well-formed**) and

"+1 Every exceeds ="

is excluded. Finally, you have to define what it means for one sentence to **follow** from others; in other words, you must establish rules of inference.

You do not have to adopt Aristotle's set of rules. (Historically, that set was viewed as the complete, authoritative description of deduction, as Euclid's geometry was held to model the universe.) You could decline to include universal instantiation (hereafter "UI"). That would eliminate our way of particularizing general statements; it would weaken, but not invalidate, your system. You could add rules, like **reasoning from the converse**. Thus, from the combination

"If a quadrilateral has four congruent sides, then it has perpendicular diagonals."
"Quadrilateral ABCD has perpendicular diagonals."
you would conclude
"Quadrilateral ABCD has four congruent sides."
The disadvantage there is that you would deduce contradictions, because the conclusion might be false. (Build an example of a quadrilateral with perpendicular diagonals but not congruent sides. Could the sides have four different lengths?)

b) Gödel's coding

[For a mathematically precise, yet readable, 118-page presentation of the following, read Ernest Nagel and James R. Newman's *Gödel's Proof*. It is now available, like **Boyer**, at archive.org.]

In 1931, the Czech-born mathematician Kurt Friedrich Gödel (1906-1974) published a method for completely representing a formal system *with numbers*. [There is no sound in English to match "ö". About the best we can do is GUH-del.] He worked entirely with symbols, but for simplicity we will allow ourselves to use words and to modify his numbering scheme.

Start with any way to number letters and symbols. For example, under the ASCII standard, all the symbols on an ordinary keyboard are numbered 32-126. (Earlier numbers are reserved for "control characters," mostly instructions to a printing device. Thus, "Form Feed", meaning go to the next sheet of paper, is numbered 12. Notice that, for example, "≅" is not on the ASCII list; for any such absent symbol, we substitute words.) Take a prime bigger than all the character numbers. Even "extended ASCII" stops at 255, so the prime 257 will serve. Starting there, list the primes:

#1. 257 #2. 263 #3. 269 #4. 271 ….

We use them to represent elements of the formal system.

> Look at the word "Every". Its letters are ASCII-numbered 69, 118, 101, 114, 121. We may represent it, in isolation, by the natural number
> $I = 257^{69}\,263^{118}\,269^{101}\,271^{114}\,277^{121}$.
> This number cannot be misinterpreted as another word's number, because prime-power factorization is unique if the primes are in increasing order. Therefore we can recover the word from I.

With such coding, we can turn statements about words into statements about numbers. For example, the statement that "Every" is a five-letter word allowed to begin a sentence (owing to the upper-case first letter) turns into

I is divisible by five primes, the smallest of them raised to a power from 65 to 90,
and the others raised to powers from 97 to 122.

> By extension, the same coding, and the same observations, apply to sentences. Thus,
> Every element has a follower
> [Starting here, we dispense with the quotation marks and the ending period.] reads in ASCII as
> 69-118-101-114-121-32-101-108-101-109-101-110-116-32-104-97-115-32-
> 97-32-102-111-108-108-111-119-101-114.
> Those 32's represent the spaces, which indicate that the sequence *is* a sentence and not some long word. We can use the first twenty-eight primes to represent the sentence by
> $J = 257^{69}\,263^{118}\ldots 409^{119}\,419^{101}\,421^{114}$.
> Then the statement that the sentence is a universal (starts with "Every") becomes simply
> J is divisible by I.

We extend the coding further to sequences of sentences. View the sequence

Every element has a follower
1 is an element

We first note that in any sequence, the sentences are separated by an invisible ⟨CarriageReturn⟩, ASCII number 13. Therefore we represent the sequence by

$$K \;=\; J \times 431^{13}\, 433^{49}\, 439^{32}\, 443^{105}\, 449^{115}\, 457^{32}\, 461^{97}\, 463^{110}\, 467^{32} \times 479^{101}\, 487^{108}\, 491^{101}\, 499^{109}\, 503^{101}\, 509^{110}\, 521^{116}.$$

The statement that K represents a two-sentence sequence turns into the number statement

K has exactly one prime raised to power 13, and the prime-power products
on either side of the one prime are in the set of sentence numbers.

Finally, we can turn the rules of inference into statements about numbers.

We agree that

Every element has a follower
1 is an element
1 has a follower

is a valid inference under the UI rule. Within the sequence, its three sentences have codes

$J, \qquad M \;=\; 433^{49}\, 439^{32} \ldots 521^{116}, \quad N \;=\; 541^{49} \ldots 631^{114}$ (figure out the rest of N).

Hence the sequence has code

$$J \times 431^{13} \times M \times 523^{13} \times N.$$

Turn the inference rule into a function *FolByUI*. Then the number statement

$$N \;=\; FolByUI\,(J, M)$$

codes the inference statement that sentence N follows by UI from J and M.

c) the significance of the coding

We see that Gödel's coding enables us to turn statements about a formal system (**metastatements**) into statements about number relations. If the system is rich enough to encompass arithmetic—for example, if its set of axioms includes the Peano axioms—then those number-relation statements can be rendered as statements *within the system*.

With that in mind, define **proof** as a sequence of sentences in which each one is an axiom in the system, or else follows from one or more preceding sentences by some rule of inference. We can check any sentence's number to verify that either it conforms to the description of some axiom's number or it is the value of some inference-function applied to previous sentence numbers. Therefore we can use numbers to state that a given sequence "proves" its last sentence. Thus, if

$$J \times 431^{13} \times M \times 523^{13} \times N$$

belongs to the set of proof numbers, then the sequence it codes proves that 1 has a follower.

(Contrast that with our notion of proof. For the base-angles theorem, we said that if AC and BC are congruent sides, then triangles ACB and BCA are congruent by SAS, making the base angles congruent. Under the definition here, we would have to write out the relevant preceding axioms—SAS was the eighteenth axiom—plus all the theorems that went into making definitions and establishing relations, plus every rule of inference applied along the way. A formal system is a laborious undertaking.)

If the formal system encompasses arithmetic, then we can make that "sequence … proves …" statement within the system. Accordingly, the system has statements that say, in effect,

Statement number N has (is the last sentence of) a proof.

It also has the denials of such statements,

Statement number n does not have (is never the last sentence in) a proof.

Gödel established that in any such system, there is one of that last type that makes its statement *about its own number*. In simpler words, this **Gödel statement** says of itself that it cannot be proved.

There is no contradiction there, unlike with a statement that says of itself that it is false. The predicate "is provable" is amenable to numbering; it is equivalent to "appears at the end of a proof". By contrast, "is false" cannot be expressed with numbers.

Now consider that the Gödel statement is either true or false. You cannot prove it is true, because if you did then you would have a proof of it, making the statement false and the system inconsistent. You cannot prove it is false, because if you did then you would falsify that the statement cannot be proved, it would be provable, and again the system would be inconsistent. Therefore if the system is consistent, then you cannot prove the statement true and cannot prove it false; the statement is **undecidable**. It is a statement whose truth or falsehood cannot be established by deduction. In any worthwhile system, there will be truths that we cannot obtain by deductive reasoning.

Section X.B. Twentieth Century Developments

1. Solutions

a) Hilbert's second problem

The result that a consistent system cannot be complete is usually called Gödel's *first* incompleteness theorem. There is a related second theorem. Gödel's 1931 paper also showed that you cannot prove within a system that the system is consistent, unless of course the system is inconsistent.

[Remember the logical peculiarity that you can prove *anything* from inconsistent premises.]

> Recall the "Gödel statement" (GS) that says of itself that it is not provable. If it is false, then the system is inconsistent. By contraposition (one of Aristotle's rules of inference), if the system is consistent, then the GS is not provable. Gödel showed that you can write the proof of the first theorem—proof that the GS exists—within the system. If you could also write within the system a proof that the system is consistent, then you would have proved within the system that the GS is unprovable. But there is no proof that the GS is unprovable.

In fewer words, you cannot prove that arithmetic is a consistent deductive system within arithmetic itself. In that limited sense, the second theorem settled Hilbert's second problem (section IX.D.4).

b) Hilbert's first problem

Gödel emigrated to the United States in 1940 and spent the remainder of his life at the Institute for Advanced Study. He continued his work in mathematical logic, and inspired others along the way.

In that same year, he proved a remarkable statement about the continuum hypothesis: You cannot prove that it is false. That is, suppose you add it to the set theory of Zermelo and Fraenkel (section IX.C.2, later extended with the Axiom of Choice, at which we barely hinted in IX.C.2b(ii)). Then you cannot arrive at a contradiction, unless the set theory was already inconsistent without the hypothesis. (Remember Klein's "relative consistency" between Euclidean and non-Euclidean geometry, from section IX.D.1c(iii).)

In 1963, the American Paul Cohen (1934-2007) proved that you cannot prove the continuum hypothesis is true. Thus, if you add the denial of the hypothesis to the axioms of set theory, the resulting system is consistent, unless set theory had a contradiction in the first place. (For his work in mathematical logic, Cohen received a 1966 Fields medal, together with Gödel's admiration; see St Andrews.)

c) joint efforts

Over the twentieth century, the process of discovery by individual researchers morphed into collaboration. With the evolution—by now, revolution—in transportation and communication, joint work on a worldwide basis is common.

(i) Fermat's last theorem

An example is the hunt that eventually chased down Fermat's conjecture. In section VIII.C.1d, we mentioned the work of Sophie Germain (French) around 1819. We could have added the indispensable work of Ernst Kummer (German; see St Andrews) in the 1840's. The last half of the twentieth century made the effort truly international: There was the breakthrough of Yutaka Taniyama (Japanese) and others around 1955; the connections Gerhard Frey (German) and Kenneth Ribet (US; both at Wikipedia®) made to the last theorem; and finally Andrew Wiles (English) and the solution in and past 1995.

That hunt inspired much mathematical activity, both back in Kummer's algebraic number theory and in the study of "elliptic curves" that Taniyama connected to "modular forms." In view of such spun-off research, it seems odd that Hilbert did not consider the Last Theorem worthy to include in his list of important problems.

(ii) classification of finite simple groups

A less-known example is the characterization of simple groups, the ones that are building blocks for all groups.

[As we did before, we use "group" to mean *finite* group. Likewise, we don't do trivial: "Subgroup" means *nontrivial* subgroup. Skip this part if you are not familiar with the long section IX.A.2b.]

With one type of exception, every group has subgroups. The exceptions are the groups of prime order (number of elements; see Exercise IX.A.2a(iii):4). Any subgroup H partitions its group G into cosets. If H is normal, then there is a natural way to operate on the cosets, forming *them* into a "quotient group." If normal subgroup H is maximal—you cannot find a normal subgroup strictly between H and G—then you cannot find any normal subgroup in the quotient group; the quotient group is simple.

> There is a rough analogy with prime factorization. Finding a maximal normal subgroup of a group is like finding a maximal factor of a natural number. The last "maximal" does not mean "biggest." A factor n of number N is **maximal** if you cannot squeeze another factor (strictly) between n and N. Thus, 28 is not the biggest factor of 140, but it is maximal. No number between 28 and 140 is both a multiple of 28 and a factor of 140. (Check the multiples of 28.)
>
> The only way n can be a maximal factor of N is if N/n is prime. (Argue why.) Analogously, H is a maximal normal subgroup of G iff the quotient group (appropriately denoted by G/H) is simple. While H and G/H do not *characterize* G, they are smaller groups that give information about G.
>
> If you write a chain of maximal normal subgroups to produce a composition series
>
> $G, \quad H_1, \quad H_2, \quad \ldots, \quad H_k, \quad \{I\},$
>
> you are building something like a maximal-factor ladder,
>
> 140, 28, 4, 2, 1,
>
> which produces the prime factorization
>
> (140/28)(28/4)(4/2)(2/1) = 5 (7) 2 (2).

This quotient-group idea says that any group is built up through a kind of multiplication of factors that cannot themselves be factored. Accordingly, a characterization of all possible simple groups would help explain the structure of all groups.

That characterization took slightly more than the twentieth century. In a prize-winning article for the Bulletin of the American Mathematical Society, Ronald Solomon gives a brilliant account of the search, from the 1892 suggestion by Ludwig Otto Hölder (much better known in analysis), through the premature victory celebration of around 1980, to the near-completion status in 2001. (The final announcement of completion came in 2004.) The article has 36 pages of text, followed by more than 130 references. It estimates that the entire proof covers some 15,000 pages. One highlight—a watershed, really—of the research is the 1963 theorem of Walter Feit and John Thompson that nonabelian simple groups have to be of even order. (Feit, born in Austria, went to the University of Chicago and worked at Yale. Thompson, born in Kansas, went to Yale and worked at Chicago. Almost forty years apart, he received both a Fields Medal and an Abel Prize.) The actual statement of the theorem is this:

All finite groups of odd order are solvable.

The proof of that eight-word statement (projecteuclid.org) occupies more than 250 pages, an entire volume of the prestigious *Pacific Journal of Mathematics*.

2. Surviving Puzzles

Hilbert's "wonderful remark" (section IX.D.4) was that if he returned in a thousand years, he would first ask whether the Riemann hypothesis had been solved. He judged its difficulty well; it remains unbroken. See Wolfram about the international work on it.

The hypothesis is at advanced level on the border between analysis and number theory. There *are* remarkably elementary unsolved puzzles in number theory. Recall Fermat's primes of the form $2^{2^n} + 1$ (section VII.A.4f(iii)) and Euclid's primes $2^n - 1$ (n prime, from perfect numbers, section III.B.4d). We still do not know how to characterize them, or even whether there are infinitely many of either kind. The same goes for **twin primes**. Those are consecutive odd numbers that are prime, like 17 and 19, or 41 and 43. (How many triplets, *three* consecutive odd numbers that are prime, are there?)

The most elementary of them all is the **Goldbach conjecture**. Christian Goldbach (1690-1764) wrote in 1742 to his friend Euler after noticing a pattern. Given a reasonable even number, Goldbach could always find two primes adding up to it. Thus,

$50 = 3 + 47$

is easy. More interesting is

98	=	3 + 95,	with 95 divisible by 5;
	=	5 + 93,	by 3;
	=	7 + 91,	by 7;
	=	11 + 87,	by 3;
	=	13 + 85,	by 5;
	=	17 + 81,	by 3;
	=	19 + 79,	both prime.

Goldbach wondered whether every even number is such a sum. The truth is still unknown. See work on it at Wolfram.

3. The Unexpected Application

The accelerating advances in communications, from wireless telephones to the partly-wired internet, have put a premium on secure signals. The signals are [relatively] easy to intercept. Their senders want them to be incomprehensible to all but the intended recipients. The most popular cipher (method for encrypting messages) is based on [relatively] elementary number theory.

a) ciphers

The most elementary ciphers simply substitute for each letter (of the appropriate alphabet) a different letter or symbol. The easiest such ciphers merely shift the alphabet. If you assign

$$A \to C, \qquad B \to D, \qquad \ldots, \qquad X \to Z, \qquad Y \to A, \qquad Z \to B,$$

then

Every element has a follower

becomes

Gxgta gngogpv jcu c hqnnqygt.

If that seems lame, you can make the encryption harder to solve. First, make all the letters uppercase and lose the spaces. The message

GXGTAGNGOGPVJCUCHQNNQYGT

has fewer hints about the structure of the sentence. (If you do that, you have to avoid writing

SPEAKINAUDIBLETONES

or else count on the recipient to know whether there should be a space after "IN".) Second, use a permutation of the alphabet, not a simple shift. Either of those has the disadvantage that every occurrence of a given letter is represented by the same letter or symbol. (The encryption yields to "frequency analysis.") Third and more effective, you can make a reasonably secure cipher by using one shift to encrypt the first letter, a second shift for the second letter, and so on. The shifts could be decided by a sequence of letters in something commonly available [like the first seventeen lines of Guzman's section X.B.3a] or—even better—by a random sequence of integers from 1 to 26 that you and your recipient both possess.

Clearly the last paragraph describes embellishments, improvements to security. At bottom, any cipher is a transformation F with two properties: It turns MESSAGES into

NONSENSE = F(MESSAGES);

and *it is invertible*. You can supply both F and F^{-1} to your recipients, but just one will do. Knowing either of them allows anyone with enough cleverness, knowledge, or tools to deduce the other. For that reason, you want to keep F and F^{-1} out of the hands of the enemy.

b) the RSA algorithm

Internet signals, just as those carried by radio waves, are accessible to anybody with the proper receiver. To enable us to blow money from the comfort of home, those signals need protection; they have to be encrypted. Given the huge number of potential customers, **public-key systems** are desirable. Those are encryption transformations that can be exposed to everybody, because the problem of finding their inverses—never impossible—is intractable.

The best known public-key system uses an algorithm developed at MIT in 1977 by Ron Rivest, Adi Shamir, and Leonard Adleman. It takes advantage of properties of Euler's totient function φ.

> All the following facts are from section VIII.C.1d.
>
> By definition, $\varphi(m)$ is the number of naturals below m that are relatively prime to m. The function is multiplicative:
>
> If m and n are relatively prime, then $\varphi(mn) = \varphi(m)\varphi(n)$.
>
> If p is prime, then all of $1, 2, \ldots, p-1$ are relatively prime to p. Therefore $\varphi(p) = p-1$. If p and q are unequal primes, then
>
> $$\varphi(pq) \;=\; \varphi(p)\varphi(q) \;=\; (p-1)(q-1).$$
>
> Finally, by Euler's generalization of Fermat's Little Theorem, if k is relatively prime to m, then
>
> $$k^{\varphi(m)} \equiv 1 \qquad \text{mod } m.$$

(i) the elements

For illustration, let us stick to the uppercase letters A-Z. In

EVERYELEMENTHASAFOLLOWER

the ASCII numbers are

69, 86, 69, 82, 89, 69, ….

We will encrypt that sequence by "triples," three digits at a time. Such grouping,

698, 669, 828, 969, …,

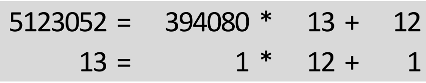
at least avoids representing every "E" by the same 69. If needed, we would append 0 or 00 to the end.

Start with two primes having more digits than the triples, say p = 1123 and q = 4567. Choose next a number K smaller than p and q and relatively prime to

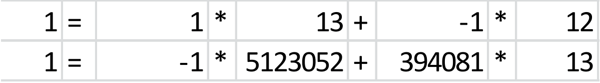
$$L = \varphi(pq) = 1122(4566) = 5\,123\,052.$$

A good choice is K = 13. (Is there a quick way to check that 13 is relatively prime to L?) That K is the encryption key. We announce it to the world, together with pq. We evaluate the inverse of K modulo L. (Why must that inverse exist?) That inverse is the decryption key; we guard it zealously.

> [I didn't pick 13 at random. I chose it because it divides L + 1:
>
> $$13 \times 394\,081 = 5\,123\,053 = L + 1$$
> $$\equiv 1 \mod L.$$
>
> That congruence says that
>
> $$K^{-1} = 394\,081 \mod L.$$
>
> Finding a divisor for L + 1 is straightforward. You can simply divide it by 2, 3, …, p – 1. I suspect a spreadsheet will do those 1121 divisions in less than
>
> $$10 \times 1121 \text{ microseconds} \cong 1/90 \text{ sec.}$$
>
> This computer-time issue will become important in a while.]
>
> A divisor of L + 1 is unnecessary. It is easy to program the Euclidean algorithm into a spreadsheet to check that a candidate is relatively prime to L. Then it is (less) easy to unwind the algorithm to write the GCD as an integer combination. For K =13, just two rows are needed:



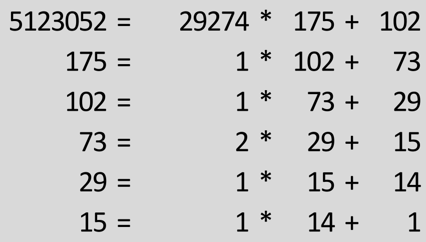
5123052 =	394080 *	13 +	12
13 =	1 *	12 +	1



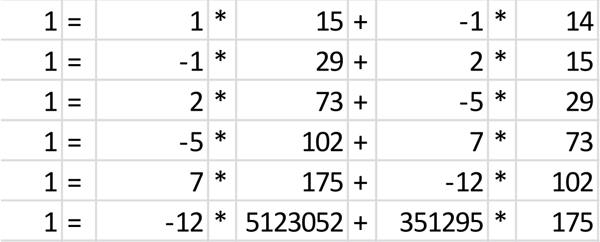
1 =	1 *	13 +	-1 *	12
1 =	-1 *	5123052 +	394081 *	13

.

> For a less trivial chosen number, it is just a matter of filling down. Thus, K_1 = 175 produces



5123052 =	29274 *	175 +	102
175 =	1 *	102 +	73
102 =	1 *	73 +	29
73 =	2 *	29 +	15
29 =	1 *	15 +	14
15 =	1 *	14 +	1



1 =	1 *	15 +	-1 *	14
1 =	-1 *	29 +	2 *	15
1 =	2 *	73 +	-5 *	29
1 =	-5 *	102 +	7 *	73
1 =	7 *	175 +	-12 *	102
1 =	-12 *	5123052 +	351295 *	175

.

> Accordingly, $K_1^{-1} = 351\,295$ modulo L.

(ii) encryption and decryption

Encryption and decryption both work modulo pq.

To encrypt the sequence of triples, the guy writing to us raises each to the power K. He will send

$$698^{13}, \; 669^{13}, \; 828^{13}, \; 969^{13}, \; \ldots.$$

He does not compute *any* of those directly.

> Our correspondent wants those modulo pq. To do the (absolutely) necessary operations, he remembers Egyptian multiplication (section II.A.2).

binary 13	power of 2	power of 698	running product
1	1	698	698
0	2	487204	698
1	4	4475395	422441
1	8	1011027	3350132

In the table at right, the third column lists, modulo *pq*, the powers

$$698^1, \quad 698^2 = (698^1)^2, \quad 698^4 = (698^2)^2, \quad \ldots.$$

Notice that each entry after the first requires one squaring and one integer division (to get the remainder mod *pq*). Each entry is 698 to the power in the second column. The first column (bottom to top) converts 13 to binary:

$$13 = 1101_{\text{binary}} = 8 + 4 + 1.$$

The fourth column accumulates a running product of those powers of 698 for which there is a 1 in the first column. Thus,

$$698^{13} = 698^1\, 698^4\, 698^8 = 698 \times 4\,475\,395 \times 1\,011\,027 = 3\,350\,132 \quad \text{mod } pq.$$

binary 13	power of 2	power of 669	running product
1	1	669	669
0	2	447561	669
1	4	2740225	2249988
1	8	2343496	4899653

Similarly, this second table gives

$$669^{13} = 4\,899\,653.$$

Our writer sends the sequence

$$3\,350\,132, \quad 4\,899\,653, \quad \ldots.$$

To decrypt the message, we raise each of those numbers to the power K^{-1}.

For such a big power, we have to turn the table on its side and miniaturize it. Take a look (magnify as necessary):

binary	394081	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1
power	of 2	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	16384	32768	65536	131072	262144
power of	3350132	3350132	1496153	4017513	1212378	4274212	5054484	705474	538036	1209033	2936456	1102571	2459552	5001276	4603478	597074	3703307	3637645	1641565	2810487
running	product	3350132	3350132	3350132	3350132	3350132	3959022	2956353	2956353	1030188	1043475	1043475	1043475	1043475	1043475	1043475	1043475	1043475	4346749	698

From the top row, we have

$$K^{-1} = 394\,081 = 262\,144 + 131\,072 + 512 + 256 + 64 + 32 + 1.$$

At the bottom right, we see the result

$$(3\,350\,132)^{1 + 32 + 64 + 256 + 512 + 131\,072 + 262\,144} \equiv 698 \qquad \text{mod } pq.$$

We recover the original triples.

To see why the decryption works, consider that our friend transformed each *n* by raising it to the power *K*, then we raised n^K to the power $M = K^{-1}$, both modulo *L*. The relation

$$KM \equiv 1 \qquad \text{mod } L$$

means that

$$KM = jL + 1 = j\,\varphi(pq) + 1 \qquad \text{for some } j.$$

By Euler's generalization,

$$(n^K)^M = (n^{\varphi(pq)})^j\, n^1 \equiv (1)^j\, n \quad \text{mod } pq.$$

Raising to the K^{-1} undoes the encryption.

(iii) the barrier

The basis for the public-key use of the RSA algorithm is that multiplication is easy and its inverse, factoring, is hard.

Encryption uses *K* and *pq*, both of which we reveal to everybody. Decryption requires φ(*pq*). We know that number, and we easily find K^{-1}. The rest of the world has to figure out

$$\varphi(pq) = (p - 1)(q - 1)$$

from *pq*. The world needs to factor *pq*.

To see how you get $\varphi(n)$ from the factorization of n, work with
$$\varphi(144) = \varphi(2^4\,3^2) = \varphi(2^4)\,\varphi(3^2).$$
Of the 2^4 naturals from 1 to 2^4, 2^3 are multiples of 2. The remaining
$$2^4 - 2^3 = 2^3\,(2 - 1)$$
naturals are not divisible by 2, are therefore relatively prime to 2. Hence
$$\varphi(2^4) = 2^3\,(2 - 1).$$
Similarly, of the integers from 1 to 3^2, 3^1 are multiples of 3. It follows that
$$\varphi(3^2) = 3^1\,(3 - 1).$$
Always, if the prime-power factorization of n is $P^aQ^bR^c\ldots$, then (as Euler knew)
$$\varphi(n) = P^{a-1}\,(P-1)\,Q^{b-1}\,(Q-1)\,R^{c-1}\,(R-1)\,\ldots.$$

To contrast multiplication and factoring, time yourself on this problem:

Find the product of the two primes 4567 and 6781.

You may use paper, a calculator, or a computer. [If you can do it in your head, go for it.] Then time yourself on this one:

Find the (two primes) whose product is 39 180 329.

Imagine trying to factor that product with a spreadsheet in our “straightforward” (or is it brute force?) way—that is, dividing by 2, 3, …. Undoubtedly, a spreadsheet that allows 10000 rows would do the $\sqrt{(39\,180\,329)} \approx 6000$ (potentially) needed divisions in less than ten times that many microseconds, roughly 0.06 sec. It would take seconds—the seconds you would spend to set up the calculations—to break our code if we based the encryption and decryption on that product.

Now imagine if we used the product of two primes of four *hundred* digits each. Even if you upgrade your spreadsheet to one that can do each division in 10^{-15} sec, it would labor for something like
$$10^{400} \times 10^{-15}\text{ sec} \cong 3.2 \times 10^{377}\text{ years}$$
to do the full set of divisions. (The universe is believed to be about 15×10^9 years old.)

You can do better than divide by all the integers. You could choose to divide just by some primes.

Recall the job of factoring 39 180 329, knowing that it has four-digit prime factors. Below 1000, there are (prime-number theorem) about
$$\Pi(1000) \approx 1000/\log_e 1000 \approx 145$$
primes. (The actual number is 168.) Below 10000, there are about
$$\Pi(10000) \approx 10000/\log_e 10000 \approx 1086,$$
around 7.5 times as many. That means there are about $6.5 \times 145 \approx 940$ four-digit primes (actual count: 1061). Assuming your computer can tell primes, you can reduce the 6000 divisions to about one-sixth that many.

Counting similarly at the 400-digit level, we find
$$\Pi(10^{400}) \approx 10^{400}/\log_e 10^{400}$$
$$= 10^{400}/(400 \log_e 10) \approx 1.1 \times 10^{397},$$
$$\Pi(10^{399}) \approx 10^{399}/(399 \log_e 10) \approx 1.1 \times 10^{396}.$$
Those imply that there are around 9.9×10^{396} primes of four hundred digits. (That means we have plenty to choose from in making our key.) You, trying to break the code, can cut your computer time to just 2×10^{364} times the age of the universe.

These are fanciful numbers. [I have seen estimates that 800-digit products can resist current (2015) factoring techniques for billions of years.] The crucial thing is that the RSA algorithm is a coding method with this property: Its encryption keys may be freely disseminated, without fear of interception, because the time needed to deduce from them the decryption keys would render encrypted information ludicrously obsolete.

4. Computers

Having let the genie out of the bottle, we will not elaborate on all the changes in everyday life that advances in computers and sensors have wrought. We will merely cast brief glances in two directions. One is the origin and nature of modern computers. The other is a specifically mathematical question they were employed to answer.

Devices to speed calculation are millennia old. Devices that calculate values, then implement decisions based on those values, go back less than two centuries. Read about the "engines" of Charles Babbage (Wikipedia®), which did arithmetic on numbers and followed instructions corresponding to the results. That is basically how the devices we now call "computers" work.

["Computer" used to refer to a human tasked with making calculations. Businesses and governments employed many such computers. Read at Smithsonian Magazine about the computers—all of them women—hired to process photographic data at Harvard Observatory, and how they ended up creating the system of star classification astronomers still use today.]

a) the nature of computing

The theoretical work behind modern computers originated with Alan Turing (English, 1913-1954; back to Wikipedia®. Turing is justly famous for a practical application of his ideas: breaking the German naval code in the opening phases of World War II. [Why was the *naval* code so important to Britain?]) His early work, in the 1930's, is grounded in the work of Gödel. Turing's 1937 paper *On Computable Numbers* ... expanded on Gödel's discoveries about algorithmic processes.

Turing gave the fundamental mathematical description of computing. He conceived the imaginary device now called a **Turing machine**. It is equipped with an infinite tape or ribbon, marked off into boxes. At appropriate instructions, the machine can wind the tape to a specified box, then based on what it finds there, move to another (or the same) box and change or keep what it finds at the latter. [If you're wondering about the infinity of boxes, versus say "4GB of RAM," just think of an actual computer that allows as much additional RAM—plus instructions for its use—as the job at hand calls for.]

b) computable jobs

To illustrate how a Turing machine would do a task, look at addition of two natural numbers in binary notation. (Babbage's engines used decimal representation. The reason for binary representation is entirely practical. It is easier to distinguish between two possible states of a component—a switch that is ON or OFF, a magnetization that is NORTH or SOUTH, a charge that is POS or NEG—than to discriminate among ten possible states.) Consider this "program":

```
Accept the two numbers and write them in boxes marked TERM1 and TERM2
Set PLACE = 1     [Our machine has internal instructions to read numbers from right to left.]
Set CARRY = 0 and SUM = (empty)
Look at the digits of TERM1 and TERM2 at binary place number PLACE
If they are both 0, replace what is now at SUM by the juxtaposition 0(SUM) and
    CARRY by the juxtaposition 0(CARRY)
    [Our machine's internal instructions direct it to do the second clause if the "If" part is true,
    to abandon this line and proceed to the next one if the "If" part is false.]
If they are both 1, replace SUM by 0(SUM) and CARRY by 1(CARRY)
If they are one 0 and one 1, replace SUM by 1(SUM) and CARRY by 0(CARRY)
If one is missing and the other is there, replace SUM by (other)(SUM) and CARRY by 0(CARRY)
If both digits are missing and CARRY $\neq$ 0, replace TERM1 by SUM,
    replace TERM2 by CARRY, and resume operating at the second line
    ["CARRY $\neq$ 0" means that the numerical value of CARRY is positive.]
If both digits are missing and CARRY = 0, report the numerical value of SUM and stop operating
    [When the machine stops processing, it tells us so; let's say it turns off the "Processing" light.]
Replace PLACE by the sum 1 + PLACE, and resume operating at the fourth line
```

Turn yourself into a machine and carry out those instructions on TERM1 = 12 = 1100_{binary} and TERM2 = 5 = 101_{binary}. The algorithm ends with the report that 12 + 5 = 010001_{binary}, an exact calculation in no more than fifteen minutes. (Why is it certain that the algorithm will end?)

[You should need about 18 written lines, doing three evaluations per line, to carry out the 12 + 5 addition. Imagine that a real computer would actually have to read 1000 instructions to do each of the needed steps. For example, the input ("Accept") of the original numbers could have all those "reads" hidden. Then the computer would be carrying out something like 54,000 instructions. If it can execute one instruction in 10^{-9} sec, then it would do the addition in a more reasonable 54 microseconds.]

More important than arithmetic is something Gödel knew, namely that what can execute algorithms can verify proofs in a formal system. Recall the definition of "proof" (section X.A.3c) as a sequence of sentences, each sentence an axiom or a consequence of earlier sentences under the rules of inference. A machine can run an appropriate algorithm to calculate a specified coding's Gödel number for any sentence sequence. It can check that the sentence numbers meet the rules for numbers of legal ("well-formed") sentences. It can check whether each sentence number is either an axiom number or the value of one of the inference-making functions applied to earlier sentence numbers. By such an algorithmic process, the machine can determine whether the sequence is a proof.

You might wonder if the machine could be asked to *concoct* proofs. Turing and others did much work on that question later. All we can give here is the usual view that producing proofs requires ingenuity, which by definition is not specifiable by algorithm.

c) the limits of computing

Turing's idea was that his machine can carry out any task describable by an algorithm (hereafter "program"). We naturally ask what jobs are out of reach of programs, and therefore beyond computers. One such job is identifying the members of a set of natural numbers.

(Notice the similarity between "jobs … out of reach of programs" and "truths … we cannot obtain by [deduction]." The connection to Gödel is clear.)

(i) computable sets

View the program below.

```
Accept the (natural) number and write it at the box labeled NUMBER
If NUMBER = 5, respond “Yes” and stop processing
If NUMBER < 5, respond “No” and stop processing
If NUMBER > 5, replace NUMBER by the difference NUMBER – 5 and resume processing
    at the second line
```

Implement the program starting with say 18, then with 25. See if you agree that the program will characterize the set of multiples of 5. It will *eventually* say “Yes” if you feed it a multiple of 5, and will eventually say “No” if you feed it a non-multiple. We call a set that some program can characterize **computable** (or **recursive**). We see that the multiples of 5 constitute a computable set. Our earlier remarks about proof verification show that the proof numbers in a formal system do likewise.

(ii) the other sets

Intuitively, it might appear that every set is computable. As we always observe, it would be ridiculous to give a name to a property if all candidates possessed it. We exhibit an uncomputable set.

In preparation, look at a more general property. Imagine replacing, in our 5-multiple program, the blue line with the line

If NUMBER < 5, resume processing at the second line.

The program now identifies the multiples of 5. That is, if you feed it a multiple of 5, then it eventually answers “Yes”. If you feed it a non-multiple, then it sits silent forever. Alternating between the blue and red lines, it never even turns off the light. A set that some program can identify (hereafter “enumerate”) this way is **recursively enumerable** (hereafter **enumerable**, which is not as ugly as **semicomputable**.)

The 5-multiple example indicates that if a set is computable, then both it and its complement are enumerable. We saw that given the 5-multiple program, modifying the blue line leaves a program that enumerates the multiples of 5. If instead we make the red line

If NUMBER = 5, replace NUMBER by the sum NUMBER + 5

and change the blue response to “Yes”, then we produce a program that enumerates non-multiples and says nothing about multiples.

The converse is also true: If a set and its complement are enumerable, then the set is computable. Suppose set *S* and complement *S** are both enumerable. That means some program *P* enumerates *S* and some program *Q* enumerates *S**. Write a program *R* that instructs as follows. [Forget now about imitating programming languages. Henceforth we write informal instruction sets that the guy in the box can understand. (One of Turing’s interesting questions was whether it is possible to tell whether the box is full of electronics or is hiding a human.)]

```
Execute the first line of program P, then the first line of program Q. If any line says “… do line
number n”, make line n the next line to be done in that program, but don’t perform line n imme-
diately. If it says “… stop processing”, abandon that program but continue the other one.
Execute the next line (if any) of program P, likewise Q.
Keep going until one of the programs directs the response “Yes”.
At that point, iff the responder was P, respond “Yes”. If the responder was Q, respond “No”.
```

Whatever number you feed *R*, either it belongs to *S* and *P* eventually speaks up, at which time *R* responds “Yes”; or it belongs to *S** and *Q* eventually speaks, at which time *R* responds “No”; *and not both*. Thus, *R* characterizes *S*, and *S* is computable.

To find an uncomputable set, we now need an enumerable set whose complement is not. Start by numbering the programs.

> Let us agree that a “program” can be any finite string of symbols, say from numbers 13 and 32-126 on the ASCII chart. Those include ⟨Space⟩ to separate words, ⟨CarriageReturn⟩ to separate lines, and ⟨Period⟩ for decimal points. Agree further that if the computer is unable to continue—say it reaches either the end of the program, or a decision you did not tell it how to handle, or a line it cannot understand, *or an order to respond “Yes”*—then it stops the processing and says so. (It turns off the “Processing” light.) Thus, we allow the three one-line programs
>
> ⟨Space⟩, $, and Stop processing.
>
> Those never produce a response, and therefore enumerate the empty set. The three-line program
>
> Accept the NUMBER
> Respond “Yes”
> Name the fifty US states
>
> enumerates the full set of naturals; but if you reverse the order of the lines, the computer immediately reaches an incomprehensible line, turns off the light, and does nothing else.
>
> We have 96 symbols. Renumber them 0 through 95. Then the finite string—the program—
>
> (symbol number m_1)(symbol number m_2)…(symbol number m_n)
>
> corresponds to a numeral
>
> (number m_1)(number m_2)…(number m_n)
>
> in base 96. That gives us a numbering of the programs.

For each natural n, let P_n denote program number n in the list and S_n the set that P_n enumerates. We recognize that P_1 never triggers a response, so that S_1 = empty set, and 1 is not in set S_1. Let

Accept the NUMBER
Respond “Yes”

be program P_k. (Observe that P_k has 32 characters, including the ⟨CarriageReturn⟩ at the end of each line. That means k exceeds $96 + 96^2 +\dots + 96^{31}$.) Then $S_k = \mathbf{N}$, wherefore k is in S_k. In other words, some naturals belong to the set they number, some do not. Let S be the set of those that belong: Thus,

$$S = \{n\colon n \in S_n\}.$$

That set is enumerable and not computable.

To enumerate S, run this program:

```
Accept a list of symbols to be stored at SYMBOL0 through SYMBOL95
Accept a natural NUMBER $m$ and represent it in base 96
Decode the base-96 numeral into the corresponding sequence of symbols
Read the sequence as a program and execute it, using $m$ as the (first) input number if the
    program asks for one.
```

Feed the computer the list of symbols and a number m. The machine will write out P_m and execute it. If $m \in S$, then by definition $m \in S_m$. The machine, running P_m, will eventually say “Yes” and stop. If instead $m \notin S$, meaning $m \notin S_m$, then P_m will never direct the machine to say “Yes.” The program enumerates S, and S is enumerable.

To show that S is not computable, it remains to show that S^* is not enumerable. Imagine that S^* were enumerable. Then some program P_K would enumerate it. What would happen if you ran P_K and input K? You would find

$$\begin{aligned} K \in S^* \quad & \text{iff} \quad P_K \text{ responds "Yes"} \\ & \text{iff} \quad K \in S_K \\ & \text{iff} \quad K \in S. \end{aligned}$$

That contradiction means no P_K enumerates S^*; S^* is not enumerable. Not every set is computable.

The influence of Cantor's diagonal argument is unmistakable. It was also there in the incompleteness theorem: Gödel adapted it to show that among the sentences

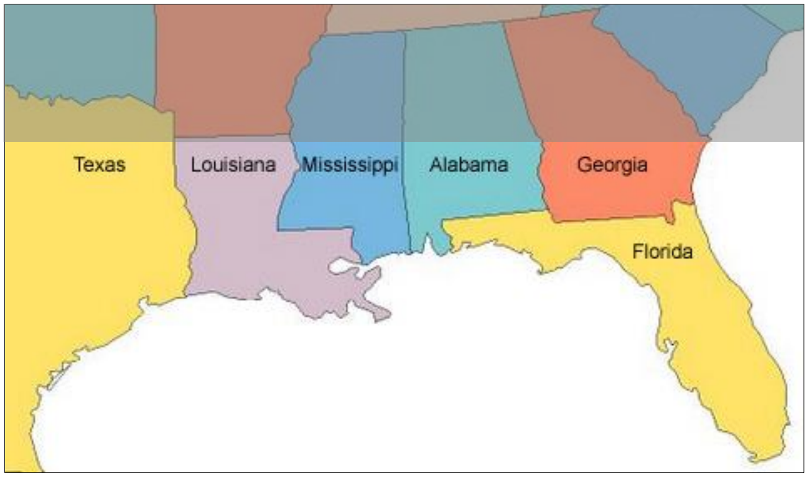
(Statement number *n* never appears at the end of a proof),

there is one that refers to its own number *n*.

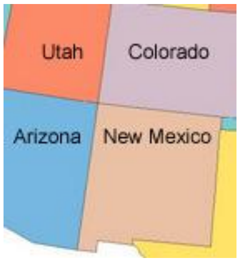
d) proof by computer

We have said that computers can check the validity of proofs in formal systems, and we suggested that they cannot engage in deductive thinking. In 1976 came an announcement of computers' taking a proof-*making* role in establishing a mathematical result, the four-color theorem.

(i) the problem



The picture at right is based on a map by the Cartographic Research Lab at University of Alabama. In that corner of the US, look at six regions: Florida, and the unshaded parts of Georgia , Alabama, Mississippi, Louisiana, and Texas. The map shows them in five different colors. There is a theorem (Wikipedia®) that five colors will always suffice to paint a map so that adjacent regions are colored differently, no matter how weirdly shaped the regions are (provided each region is contiguous. Those states are actually noncontiguous. The map ignores offshore islands.) In fact, though, those six regions need only three colors. If you make Mississippi and Texas the same orange as Georgia, and Louisiana the same blue as Alabama, then you still have different colors for regions that share a border.



["Adjacent" and "share a border" refer to regions whose boundaries have in common a curve of positive length. From the same map, the picture at left shows the place—the "Four Corners"—where Utah, Colorado, New Mexico, and Arizona adjoin. Utah shares a length of border with Colorado and a length with Arizona, but only that lone point with New Mexico. The four states are painted in four colors, but you could distinguish them with *two*. However Oklahoma (blue on and beyond the right edge) borders both Colorado and New Mexico; it needs to be a third color.]

On the other hand, return to the southeast corner. Alabama, Georgia, Florida, and the water (Atlantic Ocean on the right and Gulf of Mexico at bottom as one) all share a border with each of the other three. Accordingly, with those four regions you need four colors to distinguish regions that share a border.

In 1852, Francis Guthrie (botanist and later math professor; see St Andrews) informed the famous mathematician Augustus DeMorgan about a pattern. Working on a map of the English counties, Guthrie noted that four colors were enough to color any arrangement of regions he could invent. He wondered whether four colors were always enough. Within twenty years, lacking proof or disproof, the math world elevated Guthrie's question to the **four-color conjecture**: For any planar map of contiguous subdivisions, four colors suffice to paint it so that any regions that share a border are rendered in different colors.

(ii) the approach

Some Eulerian thinking turned the question into one about graphs. You can encapsulate the border relations via a graph with one vertex for each region and one edge connecting any pair of vertices that represent adjacent regions. Then the conjecture becomes the question whether it is possible to paint the *vertices* so that no two sharing an edge have the same color.

Over the course of a century, some simplifications were achieved. By 1976, Kenneth Appel and Wolfgang Haken had turned the question into a sort of infinite descent. They found that any graph *not* colorable by four colors had to contain at least one "configuration" from a specific set of almost 2000. (Think of four regions each bordering the other three, like Alabama-Georgia-Florida-sea, as a configuration, even though it is not among the 2000.) They later reduced the number to a mere almost 1500. They suspected these configurations made an uncolorable graph **reducible** to another uncolorable graph with fewer vertices. From that it would follow that if an uncolorable graph existed, then you could find an infinite sequence of smaller others. That would be impossible. The contradiction would show that no uncolorable map exists. All that was needed was to check that each of the 1500 configurations does imply reducibility.

The checking is algorithmic. Appel and Haken devised programs to do it. More than a thousand computer hours later, they announced their proof of the four-color *theorem*.

A big controversy arose about the claim. With traditional mathematical proof—even with such a phenomenally long project as the classification of simple groups (section X.B.1c(ii))—verification is a matter of checking for validity of inference. Such checking found reparable holes in the classification (1980) and in Andrew Wiles's original argument for Fermat (1995), and irreparable holes in an earlier argument for the four-color theorem (see Percy Heawood in the St Andrews article). For the Appel-Haken proof, verification involved checking for validity of *programming*. We saw that the European Space Agency had to incinerate a rocket (section III.B.1b) owing to an undetected programming error. NASA had a similar oversight (cnn.com). Even outfits with better-paid programmers, like Microsoft and Google, must issue the occasional corrective update. [One of my favorite principles is stated about programming, but applies to many endeavors. It is sometimes called "Murphy's Recursive Law": Debugging a program always takes twice as long as you expect, even after you apply this principle.

There are two sources you can read regarding the impossibility of error-free complex algorithms. In fiction, there is the mathematician Ian Malcolm in Michael Crichton's (book, not movie) *Jurassic Park*. In reality, there is Douglas Hofstadter's titanic masterpiece *Gödel, Escher, Bach* (yes, *that* Gödel) and the "epiphenomena" that complexity writes into your programs without your knowledge or consent.]

(iii) computerized ingenuity

By now, considerable simplification of the checking algorithms has created confidence that the computer processing really did the checking it was intended to do. But there are programs called "theorem-proving software." It is natural to wonder, as "artificial intelligence" advances, to what extent the capacity to *innovate* in a deductive system can be built into computers.

In the end, we should come back to Turing. He addressed the difference between human and machine processing in functional terms. In the **Turing test** (turing.org), you [a human, I presume] get to converse for a half hour with two agents, a human and a computer, both invisible to you. If at that point you cannot tell which is which, then you must accept that the computer was *thinking*. [The funny thing is that *computers* can sometimes tell which is which. Visit captcha.net. A **captcha** allows its home computer to do the conversing and decide whether it is talking to a human or to another machine.

The Turing test has acquired the name "imitation game," now a movie title. For some reason, the movie industry seems to have decided that people will pay to watch troubled mathematicians. Discounting the fictional one in *Good Will Hunting*, we have John Nash in *A Beautiful Mind*, Stephen Hawking (his physics is just a side effect) in *The Theory of Everything*, and Turing in *The Imitation Game*. They should consider Ramanujan's story.]

Book List

The text has multiple references to seven books.

Boyer. Carl B. Boyer, *A History of Mathematics*, Wiley, New York, 1968.
This is a classic history, whose depth and breadth of coverage are far beyond what we could hope to do here. It is now out of print, and we are fortunate to have it available online at archive.org.

Ferris. Timothy Ferris, *Coming of Age in the Milky Way*, William Morrow, New York, 1988.
Prof. Ferris sets himself the mission to explain how we arrived at our current understanding of space (the size of the universe), time (age of the universe), and creation (origin of the universe). He accomplishes his mission with a style that will captivate and enlighten anyone interested in the history of science.

al-Khalili. Jim al-Khalili, *The House of Wisdom: How Arabic Science Saved Ancient Knowledge and Gave Us the Renaissance*, The Penguin Press, New York, 2011.
Our ignorance of the achievements of Arabic scientists is scandalous. This terrific book succeeds in bringing them to our Eurocentric attention.

Kline. Morris Kline, *Mathematical Thought from Ancient to Modern Times*, Oxford University Press, New York, 1972.
This is a comprehensive and mathematically advanced history. It covers ideas in mathematics from the Babylonians through roughly where our text ends. Our references are to an online PDF version. The printed book weighed in at more than 1200 pages, nowadays broken into three volumes.

Mario Livio, *The Equation That Couldn't Be Solved: How Mathematical Genius Discovered the Language of Symmetry*, Simon and Schuster, New York, 2005.
The focus of this book is the work that settled the question of a quintic formula. It therefore has much about Abel and Galois. However, to get to them, it covers the development of algebra. That includes al-Khwarizmi, Leonardo, and the brouhahas over the solution of the cubic equation.

Merzbach. Uta C. Merzbach and Carl B. Boyer, *A History of Mathematics*, Third Edition, Wiley, Hoboken NJ, 2011.
Dr. Merzbach wrote two revised editions of Prof. Boyer's *A History* …, one in 1989-91 and a more extensive one in 2011. This later revision improves the organization. For example, the chapters on Greece have a clearer flow, separate chapters are given to India and China (for which proper names are transliterated in modern style), and there is extended coverage of twentieth-century developments. Both revisions pose a difficulty as textbooks. They appear to be intended for a general readership, and therefore lack the exercises of the original **Boyer**.

Struik. Dirk J. Struik, *A Concise History of Mathematics*, Dover, New York, 1967.
It has less than 200 pages and is correspondingly dense. It demands a pretty advanced mathematical level. However, it offers outstanding insights, especially on the history of analysis.

There are also three online resources to which the text makes multiple references.

St Andrews. Scotland's University of St Andrews [that's how they spell it] has an extensive collection of articles with both biographies of mathematicians and information about their work.

Wikipedia®. [The name is trademarked.] The articles at "The Free Encyclopedia" form a remarkable collection of information about mathematicians, mathematical exposition of their work, hyperlinked connections to related workers and work, and references.

Wolfram. Mathworld.wolfram.com has pithy (and dense) articles on a world of mathematical topics and at many levels. See their list of topics.

Finally, there is a list of books I recommended here and there. I especially favor the following five.

John and Mary Gribbin, *Annus Mirabilis: 1905, Albert Einstein, and the Theory of Relativity*, Chamberlain Bros, New York, 2005.
It gives history and description for Einstein's epochal discoveries of that year.

Robert Kanigel, *The Man Who Knew Infinity: A Life of the Genius Ramanujan*, Washington Square Press, New York, 1991.
Ramanujan's story is endlessly fascinating because the most complicated numerical relations, especially ones involving infinite series, seemed to spring full-blown into his mind. This biography captures why his (ultimately tragic) story is so interesting.

Kenneth A. Ross, *Elementary Analysis: The Theory of Calculus*, Springer, New York, 1980.
This is a great introduction to what we described as axiomatized calculus.

Dava Sobel, *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*, Walker and Company, New York, 1995.
It is a terrific account of how the problem of determining longitude at sea came down to the design of timekeepers.

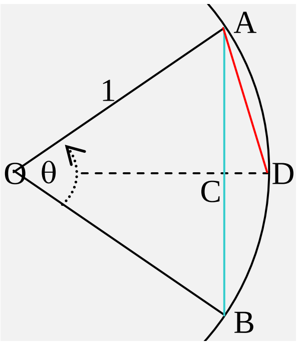
Garry Wills, *Lincoln at Gettysburg: The Words That Remade America*, Simon and Schuster, New York, 2006.
You can call Wills a philosopher on the American political system. This book is a perceptive, and eloquent, look at the eloquence of Lincoln and Lincoln's interpretation of the nation's founding document.

Index

Appendix 1: Using Chords to Measure the Circle

In Section III.A.6b(iii), we pursued Archimedes's idea to turn the geometric question of approximating π into a problem of calculation. We worked there with areas of polygons, in terms of our trigonometric functions. We noted that Archimedes had done neither: He worked with perimeters, in terms of his chord-based trigonometry. We did switch to perimeters in Exercise III.A.6b:5, but still employed modern trigonometry. Here we follow a path closer to what would have been his way.

Chord Formulas



The picture at right is from Section III.A.8a (page 43). It shows central angle AOB of measure θ in a unit circle, the angle's chord AB, its bisector OD meeting AB at C, and the chord AD of $\theta/2$. From the figure, we related the two chords by [our "half-chord formula"]

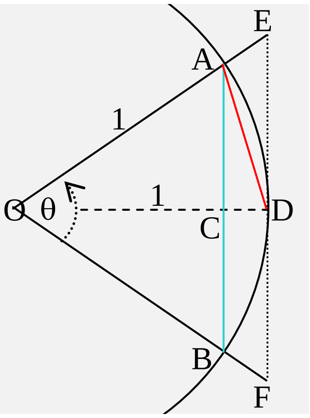
$$\mathrm{chord}(\theta/2) \;=\; \sqrt{(2 - \sqrt{[4 - \mathrm{chord}^2(\theta)]})}.$$

Check that the argument used the Pythagorean Theorem, not our ratio-based trigonometry. Also, unwind that equation to get the form

$$\mathrm{chord}(\theta) \;=\; \mathrm{chord}(\theta/2)\,\sqrt{[4 - \mathrm{chord}^2(\theta/2)]}.$$

[Feel free to do the inversion and simplification.]

A Tangent



Modify the previous figure to produce the one at right. We have added the tangent (dotted) at D, intersecting the line of OA at E and the line of OB at F. From similar triangles (Why are ED and AC parallel?), we have

$$\begin{aligned} ED/1 &= AC/OC \\ &= AC/\sqrt{[1 - AC^2]}. \end{aligned}$$

Accordingly,

$$\begin{aligned} EF &= 2\,ED \qquad\qquad \text{(Why is D the midpoint of EF?)} \\ &= 2\,AC/\sqrt{[1 - AC^2]} \\ &= \mathrm{chord}(\theta) / \sqrt{[1 - \mathrm{chord}^2(\theta)/4]}\,. \end{aligned}$$

Perimeters of Polygons

Now imagine that AB is one side of an inscribed regular n-gon, so that EF is a side of the corresponding circumscribed n-gon. The respective perimeters are

$$p_n = n\,\text{AB} = n\,\text{chord}(\theta), \qquad P_n = n\,\text{EF} = n\,\text{chord}(\theta)/\surd[1 - \text{chord}^2(\theta)/4].$$

[Here p_n and P_n are for the n-gon what P_{12i} and P_{12c} represented for the dodecagon, the case $n = 12$, in the text.] Clearly the perimeters for the two $2n$-gons are

$$p_{2n} = 2n\,\text{chord}(\theta/2), \qquad P_{2n} = 2n\,\text{chord}(\theta/2)/\surd[1 - \text{chord}^2(\theta/2)/4].$$

From the first two perimeter formulas,

$$\begin{aligned} 1/p_n + 1/P_n &= \frac{1+\sqrt{1-\text{chord}^2(\theta)/4}}{n\,\text{chord}(\theta)} \\ &= \frac{2+\sqrt{4-\text{chord}^2(\theta)}}{2n\,\text{chord}(\theta)}. \end{aligned}$$

Substituting from the half-chord formula and its unwound form, we have

$$\begin{aligned} 1/p_n + 1/P_n &= \frac{2+[2-\text{chord}^2(\theta/2)]}{2n\,\text{chord}(\theta/2)\sqrt{4-\text{chord}^2(\theta/2)}} \\ &= \frac{\sqrt{4-\text{chord}^2(\theta/2)}}{2n\,\text{chord}(\theta/2)} \\ &= 2/P_{2n}\,. \qquad \text{[Check the last two.]} \end{aligned}$$

That says

$$1/P_{2n} = 1/2\,(1/p_n + 1/P_n).$$

The perimeter of the circumscribed $2n$-gon is the harmonic mean of the perimeters of the two n-gons.

In turn,

$$\begin{aligned} p_n\,P_{2n} &= n\,\text{chord}(\theta)\;2n\,\text{chord}(\theta/2)/\surd[1 - \text{chord}^2(\theta/2)/4] \\ &= n\,\text{chord}(\theta/2)\surd[4 - \text{chord}^2(\theta/2)]\;\frac{2n\,\text{chord}(\theta/2)}{\sqrt{1-\text{chord}^2(\theta/2)/4}} \\ &= 4n^2\,\text{chord}^2(\theta/2) \\ &= (p_{2n})^2. \end{aligned}$$

The perimeter of the inscribed $2n$-gon is the geometric mean of the perimeters of the inscribed n-gon and the circumscribed $2n$-gon.

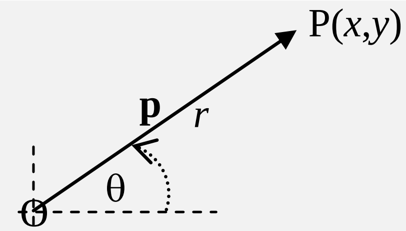
There you have the basis for Archimedes's recursion, on his terms.

Appendix 2: From Kepler to Newton

This is a calculus-based explanation of how Kepler's laws of planetary motion led to Newton's discovery of gravitation.

Preliminaries

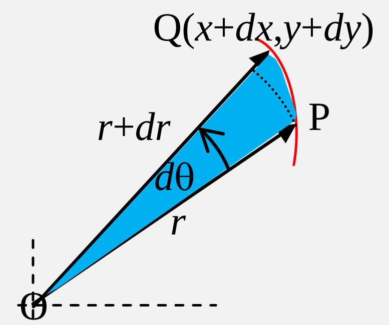
We need material from polar coordinates and vector calculus.



In the figure at right, we are in the coordinate plane. Write **p** for the vector from the origin to the point P with rectangular coordinates (x, y). Polar coordinates give a magnitude-direction description, magnitude = distance from the origin = r, direction = **azimuth** = θ radians counterclockwise around from the positive x-axis. For description in terms of components, we have

$$\begin{aligned}\mathbf{p} &= \langle x, y\rangle \\ &= \langle r\cos\theta, r\sin\theta\rangle \\ &= r\langle\cos\theta, \sin\theta\rangle.\end{aligned}$$

We will write **u** for the unit vector ⟨cos θ, sin θ⟩. (Why is it a unit vector?)



We intend **p** to indicate the position of a moving object. Along the object's path (red curve at right), consider the infinitesimal move from P to Q, coordinates

rectangular ($x + dx$, $y + dy$),

polar ($r + dr$, $\theta + d\theta$).

The position vector sweeps out the area shaded blue. That area is always approximated by the area of a sector, having central angle $d\theta$, of the origin-centered circle of radius r, namely

$$dA = 1/2\ r^2\, d\theta.$$

[You can see that the estimate misses by the area of the "triangle" upper right of the circle's arc (dotted curve), an area given by some fraction of

(length of arc) $dr = (r\, d\theta)\, dr$.

In the spirit of Barrow, we ignore that product of infinitesimals.]

Finally, in polar coordinates the unified characterization of the conic sections in terms of eccentricity leads to a single form for the equations of all four sections. Assume that our section has one focus (or if it is a circle, the center) at the origin, the point nearest the origin (necessarily a vertex if the section is not a circle) at (r_0, 0) on the positive x-axis, and eccentricity ε. Then the section is given by the polar equation

$$r = r_0(1 + \varepsilon)/(1 + \varepsilon\cos\theta).$$

Velocity and Acceleration

Now we look at position of the object as the function of time

$$\mathbf{p}(t) = r(t)\,\mathbf{u}(t).$$

The velocity is

$$\mathbf{v}(t) := d\mathbf{p}/dt = dr/dt\,\mathbf{u} + r\,d\mathbf{u}/dt.$$

[That is worth a remark: For any product of variables involving vectors, we can find derivatives by the product rule. The rule applies to a scalar multiple $f(t)\mathbf{g}(t)$, a dot product $\mathbf{f}(t)\bullet\mathbf{g}(t)$, or a cross-product $\mathbf{f}(t)\times\mathbf{g}(t)$.] We do the derivative of a vector componentwise:

$$\begin{aligned} d\mathbf{u}/dt &= \langle d/dt\cos\theta,\ d/dt\sin\theta\rangle \\ &= \langle -\sin\theta\ d\theta/dt,\ \cos\theta\ d\theta/dt\ \rangle \\ &= d\theta/dt\,\langle -\sin\theta,\ \cos\theta\rangle. \end{aligned}$$

The vector

$$\mathbf{n} := \langle -\sin\theta,\ \cos\theta\rangle$$

is also a unit vector. [Check for yourself that it is the perpendicular (“normal”) to **u** that points, from P, in the direction of increasing θ (counterclockwise).] We may therefore write

$$\mathbf{v} = dr/dt\,\mathbf{u} + r\,d\theta/dt\,\mathbf{n}.$$

[Check that result against what you would get by writing

$$d\mathbf{p}/dt = \langle dx/dt,\ dy/dt\rangle = \langle d/dt\,r\cos\theta,\ d/dt\,r\sin\theta\rangle$$

and doing those derivatives by the familiar product rule for scalar variables. Also, interpret it: The velocity has a (scalar) **radial** component (in the **u**-direction) that is simply the (signed) speed of travel away from the origin; and a **tangential** component in the perpendicular direction that would be the linear speed of an object traveling a circle of (fixed) radius *r* with *angular* speed $d\theta/dt$.]

The acceleration is then

$$\begin{aligned} \mathbf{a} &:= d\mathbf{v}/dt \\ &= d^2r/dt^2\,\mathbf{u} + dr/dt\,d\mathbf{u}/dt + dr/dt\,d\theta/dt\,\mathbf{n} + r\,d^2\theta/dt^2\,\mathbf{n} + r\,d\theta/dt\,d\mathbf{n}/dt. \end{aligned}$$

Check that

$$d\mathbf{n}/dt = d\theta/dt\,\langle -\cos\theta,\ -\sin\theta\rangle = -d\theta/dt\,\mathbf{u}.$$

That puts us at

$$\mathbf{a} = (\,d^2r/dt^2 - r\,[d\theta/dt]^2\,)\mathbf{u} + (\,2\,dr/dt\,d\theta/dt + r\,d^2\theta/dt^2\,)\mathbf{n}.$$

[Here interpretation is less suggestive, but we may still look at the elements. The d^2r/dt^2 part is unsurprising, rate of change of speed in the radial direction. The next part becomes familiar upon rewriting: Observe that

$$r\,[d\theta/dt]^2 = [r\,d\theta/dt]^2/r$$
$$= [\text{linear speed}]^2/\text{distance}$$

is Huygens's acceleration for fixed speed around a fixed circle; and it is necessarily directed radially inward, because $-r[d\theta/dt]^2$ has to be negative. (We don't do $r \leq 0$.) The tangential component has a significance, too; we will see it below.]

Kepler's Second Law

Now make our particle one of the planets, orbiting the Sun, which is stationary at the origin. As the planet moves, its location vector sweeps out area dA in infinitesimal time dt. The rate of sweep is

$$dA/dt = 1/2\; r^2\, d\theta/dt.$$

Kepler's Second Law says that this rate is constant. Hence its derivative is zero:

$$0 = d^2A/dt^2 = r\, dr/dt\; d\theta/dt + 1/2\; r^2\, d^2\theta/dt^2\,.$$

Since

$$\mathbf{a} = (\,d^2r/dt^2 - r\,[d\theta/dt]^2\,)\,\mathbf{u} + (\,2\; dr/dt\; d\theta/dt + r\, d^2\theta/dt^2\,)\,\mathbf{n}$$
$$= (\,d^2r/dt^2 - r\,[d\theta/dt]^2\,)\,\mathbf{u} + 2/r\,(d^2A/dt^2\,)\,\mathbf{n},$$

we now have

$$\mathbf{a} = (\,d^2r/dt^2 - r\,[d\theta/dt]^2\,)\,\mathbf{u}.$$

Under this Law, *the acceleration of the planets is radial* (along the line to the Sun).

[Note that

$$1/2\; r^2\, d\theta/dt = 1/2 \quad r \qquad r d\theta/dt$$
$$= 1/2\;\text{distance}\quad (\text{speed perpendicular to distance}).$$

Except for missing mass, that is half the magnitude of angular momentum. Kepler's Second Law is equivalent to conservation of angular momentum.]

Kepler's First Law

The First Law says that our planet moves along an ellipse with the Sun at one focus, our origin. Put the planet's perihelion (point closest to Sun) at $(r_0, 0)$. From

$$r = r_0\,(1 + \varepsilon)/(1 + \varepsilon\cos\theta),$$

we have

$$dr/dt = r_0\,(1 + \varepsilon)\;(-1)(1 + \varepsilon\cos\theta)^{-2}\;(-\varepsilon\sin\theta\; d\theta/dt)$$
$$= r^2/(r_0\,[1 + \varepsilon])\;\varepsilon\sin\theta\; d\theta/dt.$$

We know from the Second Law that $r^2 d\theta/dt$ is constant. Call the planet's perihelion speed v_0. At that vertex, the (tangent to the) ellipse is perpendicular to the x-axis. Therefore the angular speed at that place is

$$(d\theta/dt)_0 = v_0/r_0.$$

We conclude that the constant value of $r^2\, d\theta/dt$ is

$$r^2\, d\theta/dt \;=\; r_0^{\,2}\, v_0/r_0 \;=\; r_0\, v_0.$$

[It had not come up before, but we might as well make it explicit that we expect our planet to conform to our usual picture and orbit counterclockwise. Having

$$d\theta/dt \;=\; 2/r^2\, dA/dt$$

start, and therefore stay, positive means that also $dA/dt > 0$; it would be ugly to have area sucked up instead of swept out. Accordingly, the initial v_0 is directed upward.] Substitute into dr/dt to find

$$\begin{aligned} dr/dt \;&=\; \varepsilon \sin\theta/(r_0\,[1 + \varepsilon])\; r_0\, v_0 \\ &=\; v_0\, \varepsilon \sin\theta/(1 + \varepsilon). \end{aligned}$$

From there,

$$\begin{aligned} d^2r/dt^2 \quad &= \quad v_0/(1 + \varepsilon) \quad \varepsilon\cos\theta \quad d\theta/dt \\ &= \quad v_0/(1 + \varepsilon) \quad [r_0\,(1 + \varepsilon)/r - 1] \quad r_0\, v_0/r^2 \\ &= \quad v_0^{\,2}\, r_0^{\,2}/r^3 - v_0^{\,2}\, r_0/[\,(1 + \varepsilon)r^2\,]. \end{aligned}$$

Our last expression for acceleration becomes

$$\begin{aligned} \mathbf{a} \quad &= \quad \left(v_0^{\,2}\, r_0^{\,2}/r^3 - v_0^{\,2}\, r_0/[(1 + \varepsilon)r^2\,] - r\,[\,r_0 v_0/r^2\,]^2 \right) \mathbf{u} \\ &= \quad -\, v_0^{\,2}\, r_0/(1 + \varepsilon)\; 1/r^2\, \mathbf{u}. \end{aligned}$$

The acceleration is *inward* and proportional to inverse-square distance.

Kepler's Third Law

The acceleration's constant of proportionality $v_0^{\,2}\, r_0/(1 + \varepsilon)$ has the odd feature that it is independent of which planet we are following, but seemingly dependent on the initial circumstances and the planet's orbital eccentricity. We will see that it is the same for all planets (as indeed for anything in captive orbit around the Sun.)

Kepler's Third Law says that the square of the orbital period is as the cube of the ellipse's semimajor axis. In symbols, there is a solar-system constant K such that our planet—any planet—has period T related to its orbit's semimajor axis a by

$$T^2 = K a^3.$$

At the same time, the Fundamental Theorem of Calculus says that the area A of the ellipse is

$$\begin{aligned} A \;&=\; \int_0^T dA/dt\; dt \\ &=\; \int_0^T 1/2\; r_0 v_0\; dt \;=\; 1/2\; r_0\, v_0\, T. \end{aligned}$$

That means

$4A^2/(r_0^2 v_0^2) = T^2 = K a^3.$

Recall the geometry of the ellipse. The area is

$A = \pi ab,$

where b is the semiminor axis. The axes a and b, the focal distance c, and the eccentricity ε (figure at right below) are related by

$a^2 = b^2 + c^2, \qquad c = \varepsilon a.$

Put those together to write

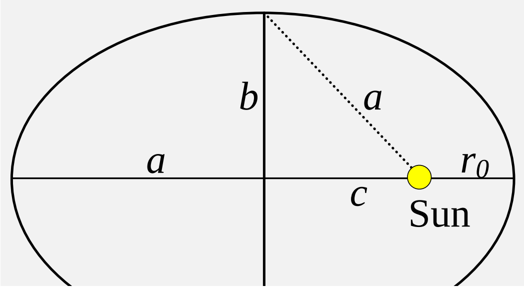
$$\begin{aligned} A^2 &= \pi^2 a^2 (a^2 - c^2) \\ &= \pi^2 a^2 (a^2 - \varepsilon^2 a^2) \\ &= \pi^2 a^4 (1 - \varepsilon^2). \end{aligned}$$

The equation for T^2 then yields

$4\pi^2 a^4 (1 - \varepsilon^2)/(r_0^2 v_0^2) = K a^3.$

Factor and rearrange to rewrite it as

$4\pi^2/K\, a(1 - \varepsilon) = r_0^2 v_0^2/(1 + \varepsilon).$



Our planet's ellipse has the Sun at one focus and perihelion r_0 away at the nearer major vertex. Accordingly,

$a = c + r_0 = a\varepsilon + r_0, \qquad$ and

$a(1 - \varepsilon) = r_0.$

Our previous equation becomes

$4\pi^2/K\, r_0 = r_0^2 v_0^2/(1 + \varepsilon).$

By Kepler's laws and Newton's calculus, all the planets—the six planets they could see—are subject to accelerations given by the unique multiple

$v_0^2 r_0/(1 + \varepsilon) = 4\pi^2/K$

of inverse-square distance.

Incorporating Newton's Laws

Now we add in Newton's laws of motion.

The First Law requires that there be a force pulling the planets toward the Sun. By the Second Law, the magnitude F of the force is proportional to the mass m of the planet:

$F = m$ (magnitude of $\mathbf{a}$) $= m\, 4\pi^2/K\, 1/r^2.$

The Third Law says that the planet exerts a like pull on the Sun. It implies that the force is proportional to the mass M of the Sun as well:

$F = G\, Mm/r^2,$

in which

$G := 4\pi^2/KM$

is a constant associated with the Sun. Thus did Newton characterize the gravitational force binding the planets to the Sun.

It was later, presumably after concluding that a single force acted on both apples near the surface of Earth and on the natural satellite a quarter-million miles away, that Newton extrapolated to a *universal* law of gravitation: Between any two bodies, there exists an attractive force given by a universal-constant multiple of the product of their masses and the squared reciprocal of their separation.

An Exercise

We have written the equivalent of

$K = 4\pi^2/MG.$

Verify that.

You can look up

$M = 1.989 \times 10^{30}$ kg and $G = 6.674 \times 10^{-11}$ m^3/kg-sec^2.

You need not look up K, because we live on a planet that has

T = one year and a = (perihelion + aphelion)/2 = 1.496×10^6 km.

Reconcile the units, then check the equality.