Precalculus

Tom Elliott

February 6, 2020

Contents

Part I

Archimedes

Chapter 1

Introduction

This book has been modified from a previous project, where it comprised the first several chapters of another project titled *Best of Calculus*. That book is here:

https://github.com/telliott99/calculus_book

I decided to make these chapters separate because the overall size of the big book makes it hard to focus on the Precalculus topics. These chapters have been the recipient of much recent attention on my part.

I wrote much of this originally as short explanations for my son Sean as he studied calculus in high school. It bothers me that so often the good stuff gets left out at that level.

(The image is a detail from a painting entitled "School of Athens", and it was used as the front cover of a wonderful book annotating the Heath translation of Euclid's

Elements).

It took a genius to figure it out the first time, but it is within anyone's grasp to appreciate what they found. I imagine myself looking over Archimedes' shoulder as he explains it to me.

There is a lot of geometry here. Most scientists I've met loved geometry in school. Proof is central to the enterprise. One of the most interesting features of this book is the natural use of proofs that I have tried to make as simple and easy to follow as possible.

My favorite authors on calculus including precalculus are Morris Kline, Richard Hamming, and Gil Strang. I highly recommend Simmons, if you can find a copy.

Finally, a saying attributed to Manaechmus (speaking to Alexander the Great), "there is no royal road to geometry". Which means, practically, learning mathematics requires that you follow the argument with pencil and paper and work out each step yourself, to your own satisfaction. That is the only way of really learning, and at heart, one of the reasons I wrote this book.

I express my sincere thanks to the authors of my favorite books, which are listed in the references and mentioned at various places in the text. Almost everything in here was appropriated from them, and styled to my taste. I offer my profound thanks also to Eugene Colosimo, S.J. He was, for me, the best of a bunch of very special teachers.

If I stole your figure off the internet, I'm sorry. I intended to redraw it but have not yet found the time.

We start with my favorite mathematician, Archimedes.

You can find the current version of this book on github here:

https://github.com/telliott99/precalculus

Chapter 2

Area of a circle

- Albert Einstein

Any fool can know. The point is to understand.

In this first unit we will develop the most famous of Archimedes geometrical contributions, a theorem on the volume of the sphere.

Before we get there we need to talk about circles (a topic to which he also contributed) and look at the volume of cones and pyramids. These are topics in geometry that come before the volume of the sphere.

area

But even before that, we need a brief introduction to area and volume. In geometry there are lines and curves and each of these has length.

Figures in the plane have area: triangles, squares and rectangles, and straight-sided polygons, so-called rectilinear figures. But also, circles and ellipses and parabolas.

Then there are solid figures, like cubes and pyramids, and cones and spheres, that have volume.

For a rectangular figure, it is easy to see why the definition of area as length times width makes sense. For a cube, the volume is the length times width times height.

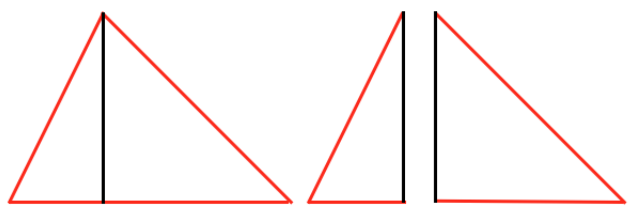
One of the miracles of calculus is that it can give us areas and volumes of curved figures. But some of those results were available from Greek geometry, even before calculus. We'll see a bit of that here. As we start reasoning about circles, we recognize that the area of a circle is going to be something *squared*, because it occupies space in the plane.

According to wikipedia

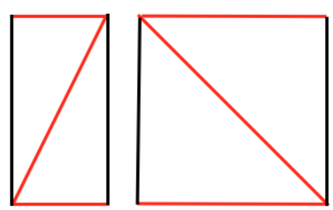
https://en.wikipedia.org/wiki/Area_of_a_circle

Eudoxus of Cnidus, born in the 5th century (408 BCE), proved that the area of a circle, like that of regular polygons, is proportional to both horizontal and vertical dimensions, and thus is proportional to the radius squared.

An idea that I bet you've run into, and we will use a lot, is that the area of a triangle is one-half the base times the height. This is easy to prove. Here we show only an acute triangle, but the theorem is true for all: any triangle can be divided into two right triangles.



The two pieces can then be pasted to their rotated equals to form two rectangles.

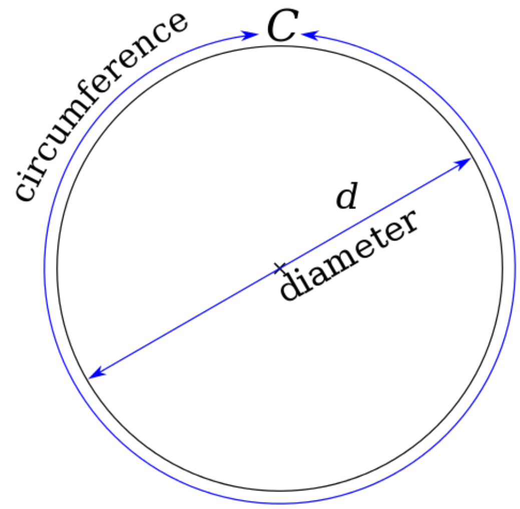


The total area of each rectangle is the total base times the height. The total area is the total base times the same height. This is twice the area of the original triangle.

□

circumference

We move on to the circle. A fundamental result about circles is that the ratio of the circumference of a circle to its diameter is independent of the size of the circle. All circles have the same shape.



The proportionality constant was named in the early 1700s and popularized by Euler a few decades later:

$$\pi = \frac{C}{d}$$

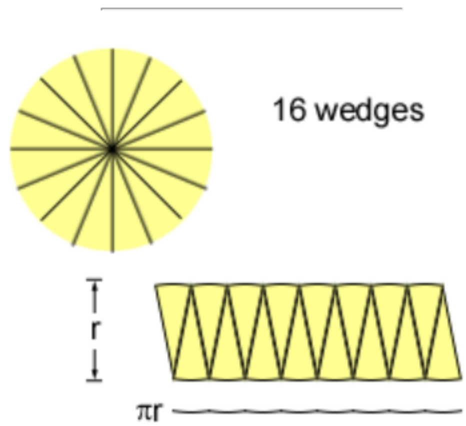
Since the radius is one-half the diameter, $2r = d$ and

$$2\pi r = C$$

This is usually stated as a self-evident fact, but it is actually a theorem to be proved. We defer that one for the moment.

area of a circle: pizza proof

Imagine dividing a circle into wedges, like you might do with a pizza. Here, the pie has been divided into 16 parts.



Since the pieces are triangular, it is easy to stack them next to each other with the bases and tips alternating, as shown. Of course the bases are not straight, but have the same curvature as the edge of the circle.

The length of the short side is the radius, r, although it is angled. The original perimeter or circumference is divided into the top and the bottom of the figure, so the length of the long side is approximately one-half the circumference and thus, with length times width, we obtain

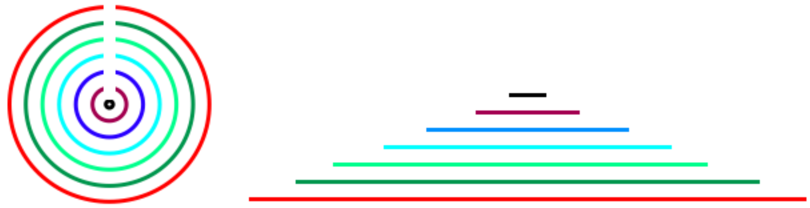
$$A = r \cdot \frac{1}{2} \cdot 2\pi r = \pi r^2$$

The trick is to imagine what happens if we subdivide the circle into many slices. The more slices, the more vertical the side, and the straighter the edges. If there are infinitely many slices, the edges will be perfectly straight and this calculation becomes exact.

The pizza proof is very much like one attributed to Leonardo da Vinci, among others.

concentric rings

Another idea is to remove concentric strips from the edge and stack them.



This is actually the same calculation as we did previously, just based on a different idea. Here, we imagine the concentric strips infinitely thin.

We obtain a triangle of height r and base $2\pi r$ so its area is

$$\frac{1}{2}\, 2\pi r \cdot r = \pi r^2$$

concentric rings

A formal proof that this triangle has the same area as the circle was given by Archimedes and is found in his *Measurement of a Circle*, proposition 1. However, many sources, including

http://www.math.tamu.edu/~dallen/masters/Greek/eudoxus.pdf

attribute the proof to Eudoxus, who was perhaps the second most famous mathematician of antiquity, and a colleague of Plato in Athens.

I love the proof, but some people have struggled with it, especially so early in the book. If you struggle, just move on. But please take a shot at it first.

We will sketch the idea and discuss it more formally **here**.

The method used is called proof by contradiction, *reductio ad absurdum*, which Hardy called

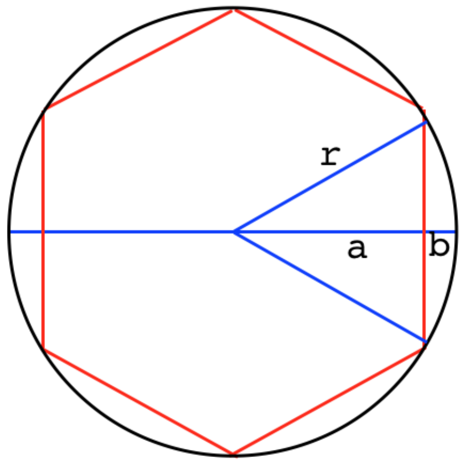
> one of the mathematicians finest weapons

One begins with an assumption. A slightly strained analogy might be, turning into a narrow street and, having missed the sign, assuming it goes your way. Later, when you meet a semi trailer head-on you know there is a problem somewhere in the logic.

inscribed polygon

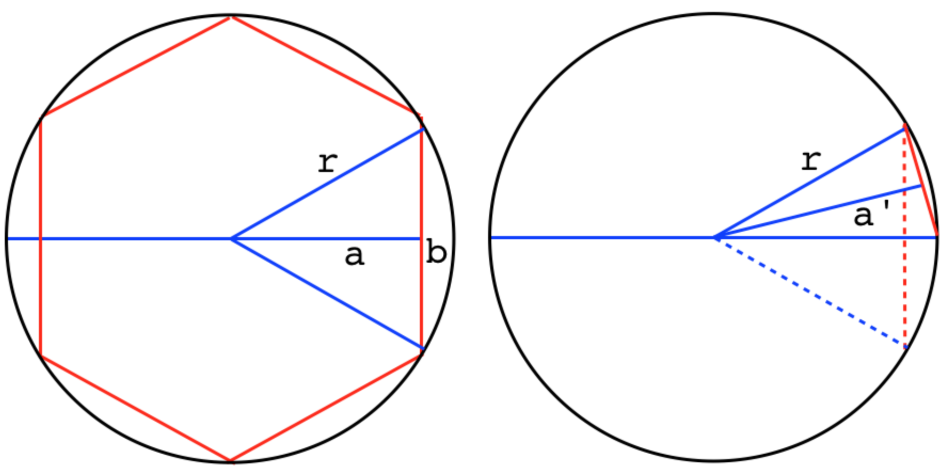
Draw a circle. Call the actual, correct, yet still unknown value for the area of the circle A.

The idea is to draw a regular polygon (all sides equal length) inside the circle.



Here we have drawn a hexagon (6-sides). If the lines from the center to the vertices have length r, then they divide the hexagon into six triangles, each with height a and base length b, where $a < r$ and b is less than the arc length. Clearly the area of the triangle is less than the area of that sector of the circle.

The base of the triangle is closest to the center (and farthest from the edge) at the center of the base, where the line segment from the center is called the apothem (labeled a). Using trigonometry, it isn't hard to calculate the area of the polygon, but we don't need to. Just call it P.



Archimedes now says, let us double the number of sides (right panel). What happens then?

The new 12-sided polygon and its component triangles will have a larger apothem a', and the total of all the bases all the way around will be larger and closer to the true circumference of the circle. There is obviously less white space between the triangle's base and the perimeter of the circle.

Thus the new P will be larger, and closer to A.

This doubling trick is easy to do. We don't even have to carry it out, just imagine doing so. We can make the difference between the area of the polygon P and the circle's true area A as as small as you please.

If your boss decides it's not close enough, just double the number of sides *again*.

That was all setup, here is the punchline.

proof

Archimedes says, let us suppose that the true area of the circle A is *not* actually equal to T (which is exactly πr^2) but is larger. Just suppose. In symbols, we are assuming that

$$T < A$$

We've already seen that $P < A$

We know we can make P as close to A as we please.

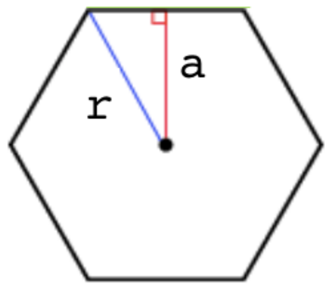
And therefore (the key point) we can make P *closer* to A than T is. The meaning of $T < A$ is that there must be some daylight between T and A. The side-doubling operation can get us into that window.

So now, by doubling, we have obtained a new P that is larger than T. We have established that

$$T < P < A$$

If $T < A$ there is really no other choice.

But look at the figure below. No matter how many sides our many-sided polygon has, $a < r$ and and the base must be less than the circumference of the circle so clearly $P < T$ for any polygon.



This is a contradiction. We have shown by two arguments which are both logically correct that $P < T$ and also that $P > T$. There is something wrong.

The resolution is that assumption we made above, that $A > T$, cannot be right. Therein lies our problem. A is not greater than T. It is either less than or equal to T.

But now try it the other way around. Circumscribe the circle with a hexagon that goes around the outside and run the argument again, and you will find that it cannot be true that $A < T$ either.

But if A is neither less than nor greater than T there is only one possibility, equality:

$$A = T = \pi r^2$$

□

Plutarch, talking about Archimedes:

> It is not possible to find in all geometry more difficult and intricate questions, or more simple and lucid explanations. Some ascribe this to his natural genius; while others think that incredible effort and toil produced these, to all appearances, easy and unlabored results. No amount of investigation of yours would succeed in attaining the proof, and yet, once seen, you immediately believe you would have discovered it; by so smooth and so rapid a path he leads you to the conclusion required.

There is one additional point. Archimedes actually provides a way of calculating the improved area of each successive polygon (or its perimeter, it is really the same problem) obtained by side-doubling.

Each cycle gives a smaller and smaller improvement, which means that there is a limiting value of this process.

That value is π, when talking about the perimeter for a unit diameter, or equivalently when talking about the area, for a unit radius. We will see how this works later, suffice it to say that the side-doubling trick gives us a way to calculate the value of π to any accuracy we have the patience to compute.

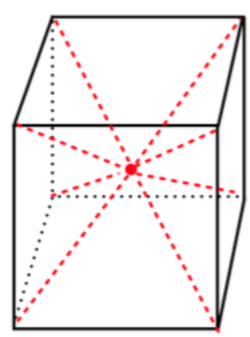
Chapter 3

Volume of a cone

We need a formula for the volume of a cone in order to find the volume of the sphere. And in order to find the cone's volume, we start with something simpler, a pyramid with a square base.

Consider a cube with all eight edges having length s. So each of the six faces is a square with sides of length s and area s^2.

Label the central point inside the solid as P. Draw lines connecting each of the 8 external vertices to P, something like this.

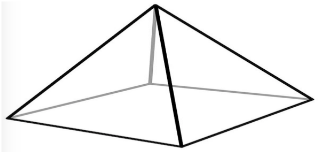


Now we imagine slicing on planes that connect adjacent pairs of lines.

You can't do this in real life by slicing up a single cube or rectangular solid, because the cuts to form one surface would ruin some of the other pieces. The cuts must enter the solid at a corner and then pivot on a line ending at the exact center.

Perhaps you could do it with a *light saber* since the beam comes to a point.

The result would be 6 identical pieces (square pyramids) looking something like this



The procedure described generates pyramids with height $s/2$. So they are a little squat, but just bear with me.

We started with a cube so that the six resulting solids would be identical.

Unfortunately you can either have six pieces come out exactly the same, as we've done, or make it so some of the pieces come out with equal base and height, but you can't do both at the same time by this construction.

Let the six identical pyramid volumes each be V, and their sum is equal to the volume that we started with. We have that

$$6V = s^3$$

$$V = \frac{1}{6}s^3$$

If we factor out the height $h = s/2$ we obtain

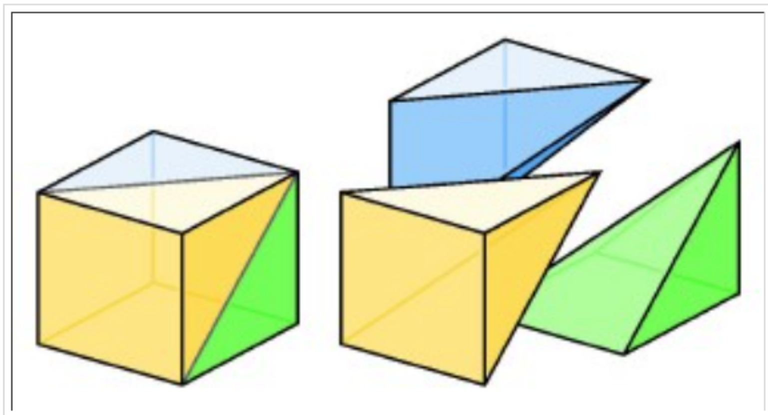
$$V = \frac{1}{3}\ hs^2$$

This is the volume for each pyramid with base area s^2 and height $s/2$.

The volume depends on both the area of the base and the height. You can show this by starting with solids that are longer in one-dimension. Since here $h = s/2$ it all works out.

better way

Here is an even better way to slice a cube

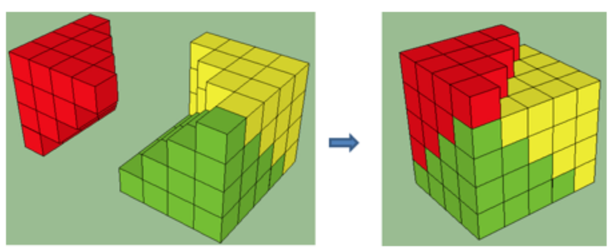


Three congruent pyramids meet along a diagonal of a cube.

When I first saw this, I thought it was a trick. But in fact, we have produced 3 identical square pyramids (they are called oblique because the apex is not in the center).

http://www.math.brown.edu/~banchoff/Beyond3d/chapter2/section02.html

Here is a version in blocks:



I know it sounds complicated but it's really not.

real world

I found a fun way to do the demonstration easily and safely. I was going to cut some wood on the table saw, but this is way better.

Get a thick piece of cheese and cut out a cube as large as you can make it and with everything squared up as accurately as you can.

Then cut straight down on a diagonal all the way through the cube, resulting in two identical pieces.

If you now take each of the pieces and orient them with the new angled surface resulting from the cut facing up, you can then make another diagonal cut straight down for each.

You will end up with a large piece and a small one. Do the same with the other half.

It seems remarkable, but the two smaller ones can be assembled into a single shape identical to each of the large pieces.

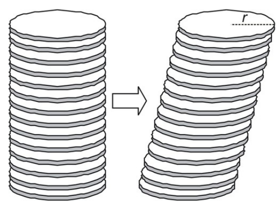
You have thus de-constructed the cube into three identical pyramids.

Good luck! It is OK to eat the demonstration afterwards.

cones

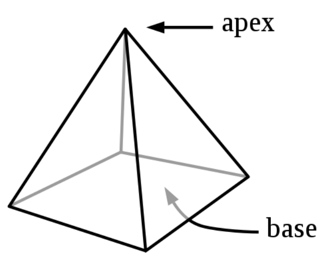
Of course, a pyramid is not a cone. But it will turn out that the volume is independent of the shape of the base. It just depends on the area. This is proved by what seem to be "hand-waving" arguments before calculus.

For example, there is Cavalieri's principle, also called the "method of indivisibles", or the "stack of quarters" argument.



If we slice a volume into small segments and then slide the slices around, the volume doesn't change. So we expect that a pyramid and a right pyramid of the same base and height have the same volume.

https://en.wikipedia.org/wiki/Cavalieri's_principle



And there doesn't seem any reason why the shape of the base should matter either, only its area. Hence we substitute a circular cross-section, i.e., a cone.

For a cone we finally obtain

$$V = \frac{1}{3}\pi r^2 h$$

Knowing basic calculus allows us to see easily where the factor of one-third comes from in the formula for the volume of a pyramid or a cone. It comes from integrating

$x^2\ dx$ and obtaining $x^3/3$. Get there we will, young padawan.

However, it is interesting to see how things might have been glimpsed in the age before calculus.

algebraic derivation of the constant 1/3

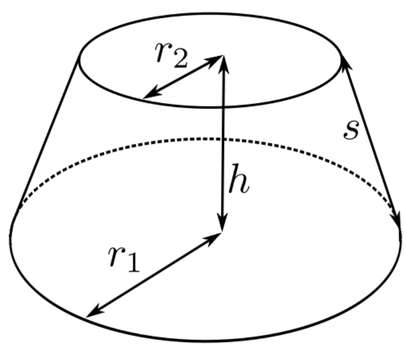
I found an algebraic argument on the web at

https://web.maths.unsw.edu.au/~mikeh/webpapers/paper47.pdf

Let us assume for this proof that the volume of a cone is proportional to both the area of the base and the height: $V = cAh$; our objective is to find the constant of proportionality.

It takes a bit of algebra to see, but gives the value for the proportionality constant as 1/3.

Consider a conical frustum, a cone with the top lopped off.



Suppose the area of the base is A and the height of the frustum is h.

Calculate the volume of the frustum as the difference between that of a larger cone with base A and height $h + e$ (e for extra), and that of a small cone (the part cut off to form the frustum) with base area a and height e.

$$V = cA(e + h) - cae$$

Now, the area of the base of a cone is π times the radius squared, and the radius is proportional to the height (depending on how sharply the side slants). Hence

$$a = ke^2$$

The area of the small base is proportional to the height squared with proportionality constant k, and the same for the large one:

$$A = k(e+h)^2$$

so we rearrange things

$$k = \frac{a}{e^2} = \frac{A}{(e+h)^2}$$

Hence

$$\frac{\sqrt{a}}{e} = \frac{\sqrt{A}}{e+h}$$

Let us manipulate this expression to find e in terms of h. It just requires a bit of facility with square roots:

$$\frac{\sqrt{A}}{\sqrt{a}} = \frac{e+h}{e} = 1 + \frac{h}{e}$$

$$\frac{h}{e} = \frac{\sqrt{A}}{\sqrt{a}} - 1 = \frac{\sqrt{A} - \sqrt{a}}{\sqrt{a}}$$

$$e = \frac{\sqrt{a}}{\sqrt{A} - \sqrt{a}} \cdot h$$

And then

$$e + h = \left[\frac{\sqrt{a}}{\sqrt{A} - \sqrt{a}} + 1 \right] h = \frac{\sqrt{A}}{\sqrt{A} - \sqrt{a}} \cdot h$$

Substituting into what we had above for the volume:

$$V = cA(e+h) - cae$$

$$= cA \left[\frac{\sqrt{A}}{\sqrt{A} - \sqrt{a}} \cdot h \right] - ca \left[\frac{\sqrt{a}}{\sqrt{A} - \sqrt{a}} \cdot h \right]$$

$$= c \left[\frac{A\sqrt{A} - a\sqrt{a}}{\sqrt{A} - \sqrt{a}} \right] h$$

This really looks like a mess.

But suppose we let $m = \sqrt{A}$ and $n = \sqrt{a}$ and then

$$\sqrt{A} - \sqrt{a} = m - n$$

then the numerator above is really just $m^3 - n^3$.

We can factor that, we get $m^3 - n^3 = (m-n)(m^2+mn+n^2)$, which you can confirm by multiplying back out. So the first term $(m - n)$ cancels the denominator. We now have:

$$V = c(m^2 + mn + n^2)h$$
$$V = c(A + \sqrt{A}\sqrt{a} + a)h$$

And, the punchline. Consider what happens as a gets larger and closer to A.

We say: let $a \to A$. But just say $a = A$.

The expression in parentheses becomes $3A$. Hence:

$$V = c(3A)h$$

But if $a = A$, the frustum has become a cylinder, whose volume we know. It is equal to Ah.

$$V = c(3A)h = Ah$$

Therefore $c = 1/3$.

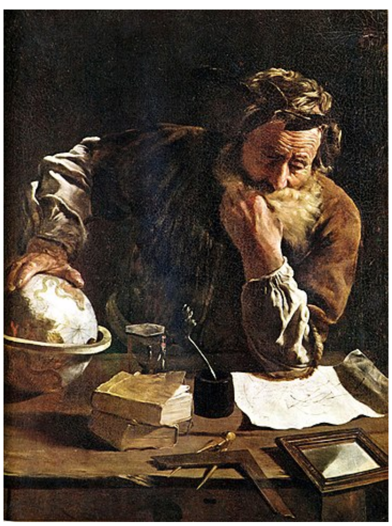
□

Chapter 4

Archimedes and the sphere

biography

Archimedes is often ranked as the greatest of the Greek mathematicians. He stands with Newton, Euler and Gauss, the best of the moderns.



Here are two images, the first is a painting by Fetti from 1620 that is in the wikipedia article on Archimedes:

https://en.wikipedia.org/wiki/Archimedes

The other is the image on the famous Fields medal, which is sometimes described as the "Nobel prize" for mathematics.

https://en.wikipedia.org/wiki/Fields_Medal

Archimedes lived and died (c.287-212 BC) in the beautiful city of Syracuse, found on the southeastern coast of modern Sicily. He is famous for many inventions, derivations and discoveries, but was evidently proud of the formula for the volume of the sphere.

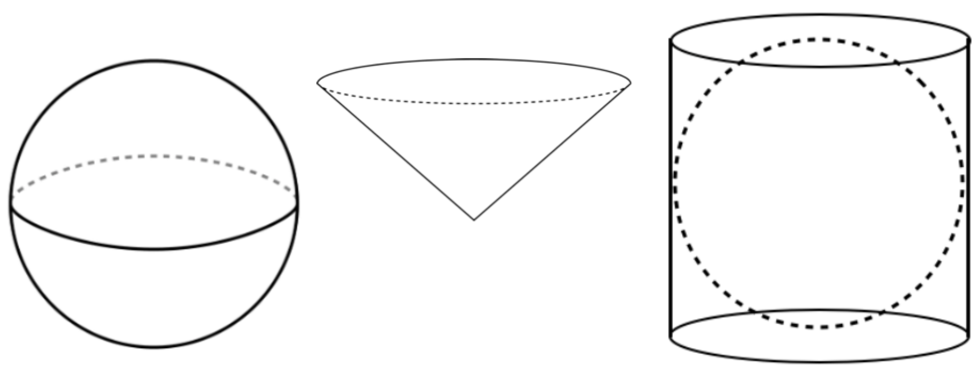
The very simple result is that the volume is two-thirds that of a cylinder that just encloses the sphere.

Because of his discovery, there was a sculpture of the sphere and cylinder carved on his gravestone, located near the Agrigentine gate of Syracuse. The grave was rediscovered by the Roman orator Cicero, covered by brush after 137 years of neglect. It is now lost again.

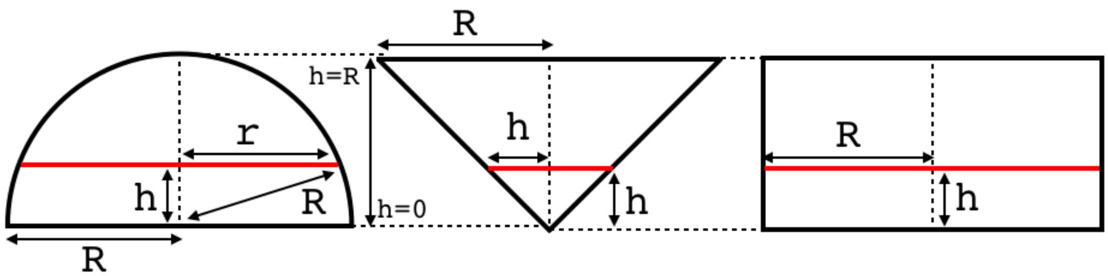
slices of solids

The following is Archimedes simple but subtle argument.

We compare a half-sphere and an inverted cone to a cylinder.



Below is a diagram showing a **vertical** cross-section through the center of each solid so we can visualize the geometry. The radius R is the same for all three. In addition, the cone and cylinder have overall height equal to R.



Now, imagine making a **horizontal** slice through each solid at an arbitrary but constant height h, shown by the red lines. I hope you can visualize each of these red slices, which are perpendicular to the page.

Each slice is a circle. Any cross-section of a sphere is a circle.

For the cylinder and cone, cross-sections perpendicular to the central axis are circles as well.

The question we ask is: what is the area for each slice?

To answer that, we need to determine the radius for each red circle.

Moving right-to-left, the radius of the cylinder is just R. For the cone, the radius at each height h is equal to h, since $R = H$. And for the sphere, we use the Pythagorean theorem to find that

$$r^2 + h^2 = R^2$$

$$r^2 = R^2 - h^2$$

For more on this theorem see **here**.

The first insight of the proof is to recognize that the radius squared for the sphere's slice (r^2), plus the radius squared for the cone (h^2) is equal to R^2, the radius squared for the cylinder.

Since the area of each circle is proportional to to the radius squared (namely $A = \pi r^2$ and so on) and

$$\pi r^2 + \pi h^2 = \pi R^2$$

so the areas add too. Our just famously and remarkable simple result: *sphere plus cone equals cylinder.*

invariance

The second crucial insight of the proof is to recognize that this property is invariant, it does not depend on which height we choose to make the slice. The three slices obtained at any height h add up like this. So if we imagine making a bunch of slices for each solid and adding them all up to find the volume, the volumes will add too.

This idea is now called Cavalieri's principle, though it was called the "method of indivisibles" before that.

The volume of the cylinder is simply πR^3. The volume of the cone is known to be one-third the area of the base times the height, or $1/3\ \pi R^3$.

Subtract to find that the area of the half-sphere is $2/3\ \pi R^3$, and therefore the volume of the whole sphere is

$$V_{\text{sphere}} = \frac{4}{3}\pi R^3$$

There is a bit of a trick here to hide the idea introduced in calculus, which makes this thinking rigorous. The sphere and cone have variable widths, which means that the radius will be different on the top of a slice compared to the bottom. Therefore, the slices have to be made very thin. In calculus they become infinitely thin, but we add up infinitely many of them.

Moreover, Archimedes was subject to certain limitations (discussed by Bressoud), which lead him to formulate the argument in terms of moments (masses of the solids and their centers). I have left that complication out of this discussion.

knowledge before proof

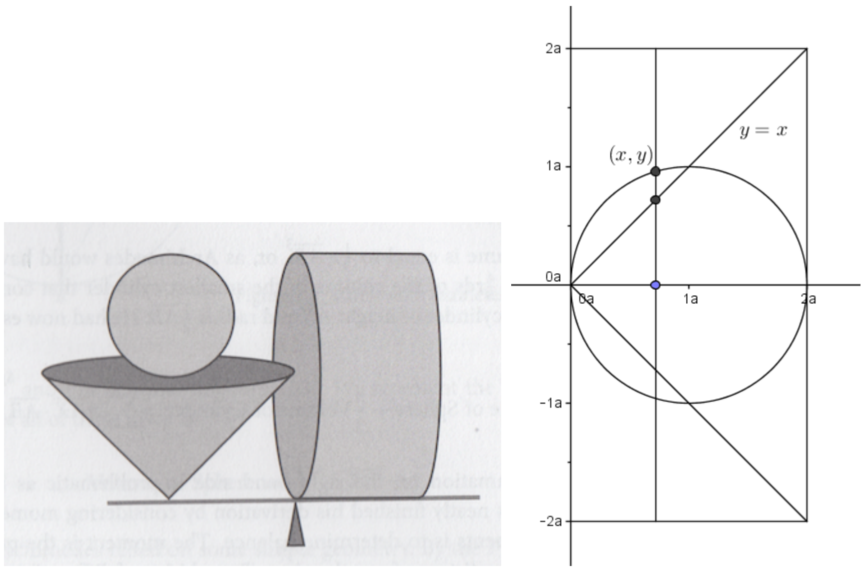
Archimedes said that he discovered the correct result by balancing the three objects on a fulcrum.

According to Archimedes (in the Method, translation by Heath)

> For certain things which first became clear to me by a mechanical method had afterward to be demonstrated by geometry...it is of course easier, when we have previously acquired by the method some knowledge of questions, to supply the proof than it is to find the proof without any previous knowledge. This is a reason why, in the case of the theorems the proof of which Eudoxus was the first to discover, namely, that the cone is a third part of the cylinder, and the pyramid a third part of the prism,

having the same base and equal height, we should give no small share of the credit to Democritus, who was the first to assert this truth...though he did not prove it.

From his description, what Archimedes actually balanced is a set-up like that shown below:



https://proofwiki.org/wiki/File:SphereVolume.png

We have a:

- sphere with radius R
- cone with radius $2R$ and height $2R$

Their combined volumes:

$$V = \frac{4}{3}\,\pi R^3 + \frac{1}{3}\,\pi(2R)^2 \cdot 2R$$

$$= \frac{12}{3}\,\pi R^3 = 4\pi R^3$$

balanced against

- a cylinder with radius $2R$ and height $2R$ with

$$V = \pi(2R)^2 \cdot 2R = 8\pi R^3$$

However, the moment of the cylinder is at one-half the distance from the fulcrum as that of the sphere-cone combination, so it balances.

Part II

Numbers and proof

Chapter 5

Integers

Integers

The *natural* or counting numbers which everyone learns very early in life are $1, 2, 3$ and so on.

One can get hung up on the question of whether the natural numbers would exist without the problem of counting a dozen sheep or all twenty of our fingers and toes. Leopold Kronecker famously said "God made the integers; all else is man's handiwork".

We will not worry about where they come from.

Mathematicians refer to the *set* of natural numbers and give that set a special symbol, $\mathbb{N}$. We write

$$\mathbb{N} = \{1, 2, 3 \dots\}$$

The brackets contain between them the elements or members of the set. The dots mean that this sequence continues forever.

How can we decide whether a particular n is in the set if we can't enumerate all of its members? We can tell by its form whether some n is a natural number or not.

If this seems problematic, you might call $\mathbb{N}$ a class instead (Hamming); we carry out *classification* to decide whether n is a natural number.

The notion of an unending sequence can be unnerving upon first encounter.

construction of N

To construct the set $\mathbb{N}$, start with the smallest element, 1. Then

$$1 + 1 = 2$$

$$2 + 1 = 3$$

$$3 + 1 = 4$$

$$\ldots$$

Add successive elements by forming $a_n + 1 = a_{n+1}$.

$\mathbb{N}$ is an infinite set.

We say there is no largest number in $\mathbb{N}$, no largest $n \in \mathbb{N}$. The symbol $\in$ means "in the set" or "is a member of the set".

Proof:

Suppose $\mathbb{N}$ did have a largest member, M.

Well, what about $M + 1$? By the definition we can construct it and it is clearly a member of the set, but $M + 1 > M$ so M is not the largest number in the set.

This is a proof by contradiction that $\mathbb{N}$ is infinite.

$\square$

set membership

Sometimes people say that

$$0 \in \mathbb{N}$$

(0 is a part of the set) but most do not, and we will follow the definition given above. If you wanted to be explicit about this you could write

$$0 \notin \mathbb{N}$$

What do we mean by infinity? We mean an upper bound on the natural numbers, and later, all rational and indeed all real numbers.

All numbers $n \in \mathbb{N}$ have the property that n is contained in the interval $[1..\infty)$. However, ∞ is *not* considered part of the interval, and that is the meaning of the the right parenthesis.

∞ is not a number so it probably doesn't even make sense to write $\infty \notin \mathbb{N}$.

least element

$\mathbb{N}$ does not have a greatest number, but it does have a smallest or least one. If pairwise comparisons are carried out, a single element, the number 1, has the property that $1 \leq n$ for all numbers $n \in \mathbb{N}$. As we go on, we will find that other types of numbers (rationals and real numbers), do not have a least positive number.

well-ordered property

Since we can also find the least member of the set excluding 1, written $\mathbb{N} \setminus 1$, we can order every number in $\mathbb{N}$.

This property is called the **well-ordered** property.

the Integers

The set $\mathbb{Z}$ contains all the members of $\mathbb{N}$ plus their negatives, as well as the special number 0, often called the additive identity since $0 + n = n$ for all $n \in \mathbb{N}$.

$$\mathbb{Z} = \{\cdots - 2, -1, 0, 1, 2, \dots\}$$

$\mathbb{Z}$ stands for the German word *Zahlen*, number. The set $\mathbb{Z}$ is usually referred to as the integers.

$\mathbb{Z}$ is also an infinite set and also has the well-ordered property. To show this simply order all numbers $n > 0$ with respect to zero using $<$, and all the numbers $n < 0$ using $>$.

Chapter 6

Primes

prime numbers

As you know, the positive integers larger than 1 are of two types:

• a prime number p has only two factors, p itself and 1

• a composite number has at least one additional factor. Either the number is a perfect square of a prime, or it has an even number of additional factors: $a_1 \dots a_k$.

The first ten primes are:

```
2 3 5 7 11 13 17 19 23 29 ...
```

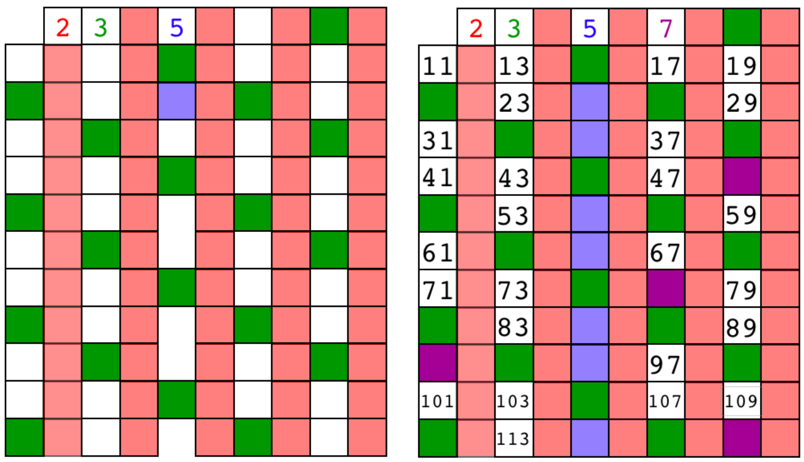
The sieve of Eratosthenes

Eratosthenes is famous in mathematics for his "sieve" which allows one to determine which numbers are prime in an economical fashion.

We will take note of him again in talking about the circumference of the earth. He was a contemporary of Archimedes and became the chief librarian at the Library of Alexandria when he was only about 35 years old.

The sieve operates by first writing down all the integers to some upper limit (here 120). To carry out the process manually it is convenient to use rows with 10 values, so there are 12 rows in all here. Most of the boxes have not yet been numbered (below, left).

Starting with the first prime number, 2, eliminate all the numbers divisible by 2 (all the red numbers, or even numbers). Here this has been done by coloring red all squares with numbers ending in $2, 4, 6, 8, 0$.



Next, do the same thing with 3 (green). 6 was already eliminated previously, but odd multiples of 3 like 9, 15 and 21 go away at this step.

The next larger number that still has a white square is 5. All the squares eliminated at this step are white ones in the fifth row, starting with 25. Continue with 7, eliminating $49, 77, 91$ and 119.

The sieve now ends (for this upper bound of 120).

The rule is that if the number for the next round, the smallest number not yet eliminated, is larger than the square root of the upper limit ($\sqrt{120}$), we terminate.

So 7 is the last value used, because after that round the smallest remaining integer is 11, but we terminate since $11^2 = 121 > 120$.

The graphic shows all the numbers which have yet to be eliminated after the round of 7. All of these numbers, 11, 13, 17, and so on, as well as those used as divisors for each round of the sieve $(2, 3, 5, 7)$, are prime numbers.

By testing for division by $2, 3, 5$ and 7, we have found the first 30 prime numbers.

2	3	5	7	11	13	17	19	23	29
31	37	41	43	47	53	59	61	67	71
73	79	83	89	97	101	103	107	109	113

From a performance standpoint, it is important that we do not need to carry out division. All that is really needed is repeated addition. Coding this algorithm in, say, Python is a good challenge.

A bigger challenge is to come up with a method to *grow* the list of primes on demand. This can be done by keeping track of the first value to be tested above the limit, for each prime in the current list.

recognizing primes quickly

There are school problems that require you to factor numbers at least up to 100, maybe more, quickly. It can be helpful to learn to recognize the primes in this range.

- First, primes end in one of the digits: $1, 3, 7, 9$.
- Test quickly for 3 as a factor by the digit sum trick.
- Test 7 as a factor by trial multiplication.

Let's do the first row, for an example:

```
11 21 31 41 51 61 71 81 91 101 111
```

You should recognize 11 as prime, immediately. Then remove the numbers whose digits add up to 3 or a multiple, leaving

```
31 41 61 71 91 101
```

Then, trial multiplication by 7 to get a number ending in 1.

The 1 reminds me of 21, so I guess: $7 \cdot (10 + 3) = 70 + 21 = 91$. Gone.

- $91 - 71 = 20$
- $91 - 61 = 30$
- $101 - 91 = 30$
- $91 - 41 = 50$

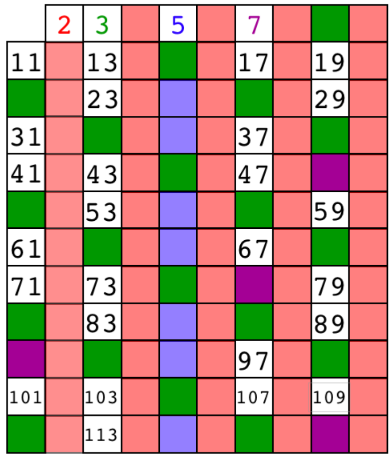
None of these differences is divisible by 7, so the numbers are not, either.

I trust you recognize that 31 is 3 more than $7 \cdot 4$.

We do not need to test any primes larger than 11. All multiples of 11 are double digits, until 110.

That leaves:

`31 41 61 71 101`



infinite primes

Euclid has a theorem and a proof that the number of primes is infinite.

Proof:

By contradiction.

Suppose the set of primes is finite, and that p_1, $p_2 \dots p_k$ are all of the primes. Construct the following numbers:

$$P = (p_1 \cdot p_2 \cdot \dots p_k)$$

$$Q = P + 1$$

For a prime number p to evenly divide Q, it must divide the difference between Q and P. But that difference is 1 and so can't be divided evenly by any prime.

Therefore, none of the known primes divides Q and at least one of these is true:

- Q is a prime not in the set of known primes
- the set was originally incomplete

The assumption that the set of primes is finite leads to a contradiction.

□

Even for a relatively small number of primes, we may encounter the second situation. Start with the first prime: 2:

$$2 + 1 = 3 \text{ (prime)}$$

$$2 \cdot 3 + 1 = 7 \text{ (prime)}$$

$$2 \cdot 3 \cdot 7 + 1 = 43 \text{ (prime)}$$

$$2 \cdot 3 \cdot 7 \cdot 43 + 1 = 1807$$

1807 is *not* prime. ($1807 = 13 \cdot 139$).

testing primality

This is a pretty deep subject. However, a simple filter to apply first is to ask: is the last digit one of $\{0, 2, 4, 6, 8\}$, i.e. is the number even? Does the number end in 5? Or is the sum of the digits divisible by 3 or 9?

There is a trick for the last test. If a number is divisible by 3 its digits add to a multiple of three. Suppose the number is:

$$abcd = a \cdot 10^3 + b \cdot 10^2 + c \cdot 10^1 + d$$

$$= a(9 \cdot 10^2 + 1) + b(9 \cdot 10^1 + 1) + c(9 \cdot 10 + 1) + d$$

If $x|(y + z)$ and $x|y$ then it must be that $x|z$. Since 9 times anything is divisible by 3, it follows that 3 must divide $a + b + c + d$ for $abcd$ to be divisible by 3.

A similar thing is true of 9 except that the digits must add only to 9.

A more general observation is that all primes greater than 3 are of the form $4k+1$ or $4k+3$, for integer k. That's because $4k$ and $4k+2$ are even, and $4k+4 = 4(k+1)$.

Any composite number n has a unique prime factorization. Its smallest prime factor p has the property (easily proved):

$$p^2 \leq n$$

Therefore, it suffices to check whether the prime numbers less than or equal to the square root of n divide n. If the square root is not an integer, we need check only the next smallest integer, what is called the *floor* of the value. If no prime less than that divides n, then n is a prime.

This can be improved still more.

https://en.wikipedia.org/wiki/Primality_test

Chapter 7

Prime factorization

We will prove that every integer has a unique *prime factorization.*

Examples:

$$6006 = 2 \cdot 3 \cdot 7 \cdot 11 \cdot 13$$

$$144 = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 3 \cdot 3$$

$$12 = 2 \cdot 2 \cdot 3$$

This is also called *the fundamental theorem of arithmetic.*

$$n = p_1 \cdot p_2 \ldots p_k$$

In the list of the prime factors of n, a factor may be repeated. To compare two factorizations for uniqueness, we suppose they are sorted (say, from smallest to greatest).

Example:

Let's try 123456789.

At first it seems easy. I get two factors of 3, leaving 13717421.

Then my luck ran out. The smallest prime factor was too large for me to find by hand. I used Python:

```
def f(n):
    for i in range(2,int(n**0.5)):
        if n % i == 0:
```

```
            print i, n/i
```

Running this

```
>>> f(13717421)
3607 3803
```

These are the two prime factors of that number.

background

We will prove that the unique prime factorization theorem is true.

Before we start on that, remember that when we say that one integer evenly *divides* another one, written as $a|n$ or a is a factor of n, what we mean is that we can find another integer k such that

$$a \cdot k = n$$

a times k is exactly equal to n.

And if there is also a number m where a evenly divides m, we write $a|m$ and mean that

$$a \cdot j = m$$

So then addition or subtraction of $m + n$ gives

$$m + n = a \cdot j + a \cdot k = a(k + j)$$

$$m - n = a \cdot j - a \cdot k = a(k - j)$$

If $a|m$ and $a|n$, a divides their sum or difference.

abnormal numbers

Hardy and Wright (*Theory of Numbers*, sect. 2:11) have a proof of prime factorization, different than the standard one, which I find quite elegant.

Its greatest virtue is that it simplifies some other proofs including of Euclid's lemma, which is tedious.

Proof.

By contradiction.

Hardy:

> Let us call numbers which can be factored into primes in more than one way, *abnormal*, and let n be the smallest abnormal number.

Start by supposing that there are two different factorizations of n:

$$n = p_1 \cdot p_2 \ldots p_k$$

and

$$n = q_1 \cdot q_2 \ldots q_j$$

where the p's and q's are all primes.

Different factorizations

As a preliminary result, consider the possibility that some p is equal to a q.

Let us rearrange if necessary so that factor is first: $p_1 = q_1$ and

$$n = p_1 \cdot p_2 \ldots p_k$$

$$n = p_1 \cdot q_2 \ldots q_j$$

But now, n/p_1 is abnormal, because it has two different prime factorizations.

That is impossible, because n is the smallest abnormal number.

Thus, we have that no p is a q and no q is a p.

If there do exist abnormal numbers with two factorizations, those factorizations must be completely different.

inequality

In this part, we establish that $p_1 \cdot q_1 < n$.

We may take p_1 to be the least p and q_1 to be the least q.

Since n is composite, either $p_1^2 = n$ (and p_1 is the only factor of n),

Or p_1 times the largest $p_k = n$.

In the second case, $p_1 < p_k$, $p_1 \cdot p_k = n$ and so $p_1 \cdot p_1 < n$.

A similar result holds for q_1.

But, since $p_1 \neq q_1$, only one of p_1 or q_1 at most, can be squared to give n. Either $p_1 \cdot p_1 < n$ or $q_1 \cdot q_1 < n$ or if there is more than one p and more than one q, both statements are true.

From this it follows that $p_1 \cdot q_1 < n$.

the contradiction

Let $N = n - p_1 q_1$.

We have $0 < N < n$ and also that N is not abnormal.

We're given that $p_1 | n$ and so, from the above equality $N = n - p_1 q_1$ and our preliminary reminder about what factorization means, it must be that $p_1 | N$.

A similar result is true for q_1, namely $q_1 | N$.

Hence both p_1 and q_1 appear in the unique factorizations of both N and $p_1 q_1$.

We have that $N = n - p_1 q_1$ and we've shown that $p_1 q_1 | N$ and certainly $p_1 q_1 | p_1 q_1$. It follows that $p_1 q_1 | n$ and hence $q_1 = n / p_1$.

But n/p_1 is less than n and has the unique prime factorization $p_2 \cdot p_3 \ldots p_k$.

Since q_1 is not a p, this is impossible.

Hence there cannot be any abnormal numbers.

$\square$

This is also called the *fundamental theorem of arithmetic*.

Chapter 8

Induction

the problem

Suppose we have some theorem that we think *might be true* for all numbers n, because we've tried it on a few different values of n and the theorem is true for all of them.

A classic example (Courant and Robbins) is this prime number generator:

$$p(n) = n^2 - n + 41$$

The remarkable function $p(n)$ produces a prime number for integer $0 < n < 41$.

```
  41   43   47   53   61   71   83   97
 113  131  151  173  197  223  251  281
 313  347  383  421  461  503  547  593
 641  691  743  797  853  911  971 1033
1097 1163 1231 1301 1373 1447 1523 1601
1681
```

But, for $n = 41$, the last two terms cancel in

$$p(n) = n^2 - n + 41$$

and then n^2 is divisible by n, thus the result cannot be prime.

By testing them all, I found that 41 is the largest prime smaller than 2000 with this property (I don't know of a proof that no more exist). The primes with this property are:

```
 2  3  5 11 17 41
```

Hamming has some other examples of theorems with many true candidates, but which are false. Here is one:

$$f(n) = n(n-1)(n-2)\dots(n-k)$$

$f(n) = 0$ for all $0 \le n \le k$, but will never be zero for any other $n > k$.

That is because there are only k zeroes of a kth degree polynomial. (As an aside, this is a consequence of the *fundamental theorem of algebra*).

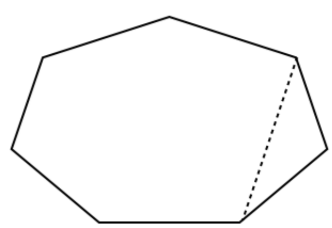
By choosing k large and expanding the definition, we can generate as many true cases as you have patience for.

Furthermore, for any function $g(n)$, $f(n) + g(n)$ will have the same property.

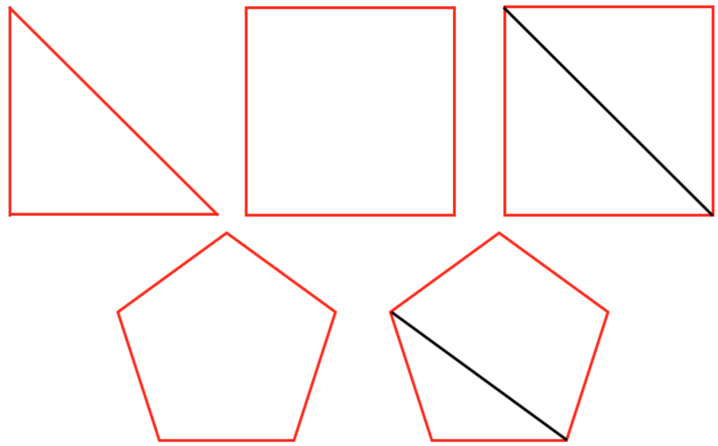
induction in geometry

In the figure below is a polygon—an irregular heptagon. Actually, there are three polygons altogether, there is the heptagon with $n+1$ sides, the hexagon with only n sides that would result from cutting along the dotted line, and the triangle that is cut off.

We want to find a formula for the sum of the internal angles that depends only on the number of sides or vertices.



The first part of the answer is to guess.



We know that for a triangle ($n = 3$), the sum of the angles is $180°$, and the sum does not depend on whether the triangle is acute, right or obtuse.

Continuing with the square ($n = 4$), we can draw the diagonal and observe that the sum of all the angles is twice $180°$ or $360°$. The partition into two triangles can be carried out with any quadrilateral, it does not require any sides being equal.

From this we guess that the formula may be:

$$S_n = (n - 2) \cdot 180$$

And indeed, in going from $n = 4$ to $n = 5$ sides we can think of the pentagon as being a quadrilateral with an extra triangle.

And in the first figure, you can see that by adding the extra vertex to go to the $n+1$-gon, we added a triangle, or perhaps you'd rather say than in going from $n+1$ to n we lost a triangle.

In all cases, the difference between n and $n + 1$ is $180°$.

The formula *seems* to work.

We can use induction to *prove* that it is correct.

The proof has two parts. We must verify the formula for a base case like the triangle, which we've done. You may wish to check that it works for the square as well, but that's not strictly necessary.

The second part of the proof is to verify that in going from n to $n+1$, we add another $180°$. The formula for n sides is $(n-2)180°$, adding another triangle gives:

$$(n-2)180° + 180°$$

That must be equal to what the formula gives for $n+1$ sides:

$$((n+1)-2)180°$$

Substituting x for $180°$ and equating the two, we have

$$(n-2)x + x = ((n+1)-2)x$$

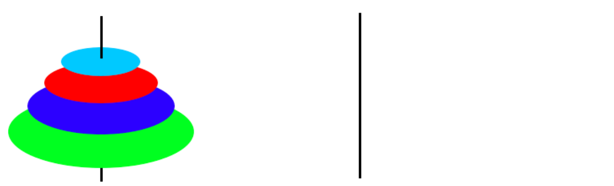
$$n - 2 + 1 = n + 1 - 2$$

$$n = n$$

which is certainly correct.

$\square$

Towers of Hanoi



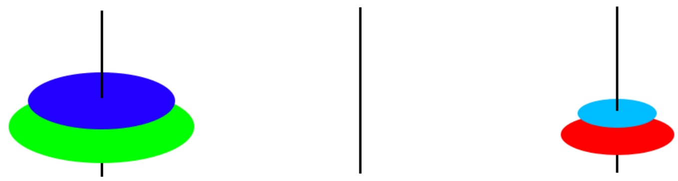
In this famous game the goal is to move a set of disks from one peg to another. Let us choose the one on the right as the target.

https://en.wikipedia.org/wiki/Tower_of_Hanoi

The rules are:

- Only one disk may be moved at a time.
- Each move consists of taking the upper disk from one of the pegs and sliding it onto another peg, on top of the other disks that may already be present on that peg.
- No disk may be placed on top of a smaller disk.

Here is an intermediate stage of the game:



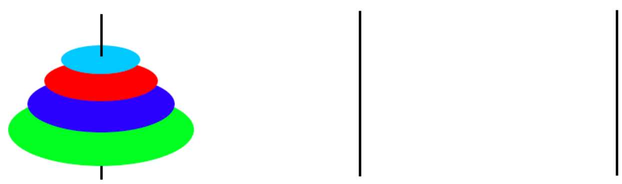
The next move is to place the blue disk on the middle peg. I think you can take it from there.

We can solve the puzzle for any number of disks n.

Proof:

By induction.

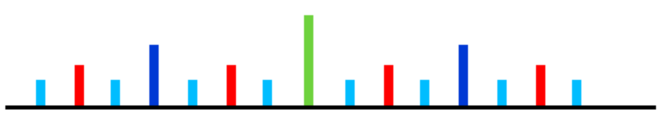
Start from the first position:



Suppose we know how to move $n-1$ disks from one peg to another. Move them to the middle peg, then move the nth disk to the right peg, then place all the $n-1$ disks on top. We have moved n disks.

The base case is to move the single light blue disk. That's trivial. The only thing to watch is if the number of disks is even or odd. If even, choose peg 2, otherwise peg 3.

Which peg is to be moved at each stage is shown in this graphic:



> The puzzle was invented by the French mathematician Édouard Lucas in 1883. There is a legend about a Vietnamese temple which contains a large room with three time-worn posts in it surrounded by 64 golden disks. The monks of Hanoi, acting out the command of an ancient prophecy, have been moving these disks, in accordance with the rules of the puzzle, since

> that time. The puzzle is therefore also known as the Tower of Brahma puzzle. According to the legend, when the last move of the puzzle is completed, the world will end.

summary

We can visualize an inductive proof as a kind of chain. We show that the base case is true, for some value of n. Then we show that if the formula works for n , it must work for $n+1$.

> Mathematical induction proves that we can climb as high as we like on a ladder, by proving that we can climb onto the bottom rung (the basis) and that from each rung we can climb up to the next one (the step).

- Graham, Knuth and Patashnik

[There is a variant called *strong* induction where we know some statement is true for *all* $0 < k \leq n$ and use it to prove something about $n+1$.]

A few more examples:

sum of digits and divisibility

It is very easy to check whether any number n is divisible by 9. Simply add up all the digits of say, 234783738:

$$2+3+4+7+8+3+7+3+8$$

$$= 5+1+1+1+1+1+0+8$$

$$= 1+8 = 9$$

Yes, 234783738 is a multiple of 9.

We propose that

$9|(10^n - 1)$ for all integers $n \geq 0$.

The statement $9|n$ means "9 divides n".

Suppose we know that $9|10^k - 1$ for some n. We mean that

$$10^k - 1 = 9x$$

for some x. Multiply by 10:

$$10 \cdot (10^k - 1) = 10 \cdot 9x$$

$$10^{k+1} - 10 = 9 \cdot 10x$$

$$10^{k+1} - 1 = 9 \cdot 10x + 9 = 9(10x + 1)$$

The right-hand side is clearly divisible by 9, and then so is the left-hand side.

The base case is $9|0$ which is true by definition but may be confusing. Try $n = 1$, then $9|(10-1)$ is certainly correct.

□

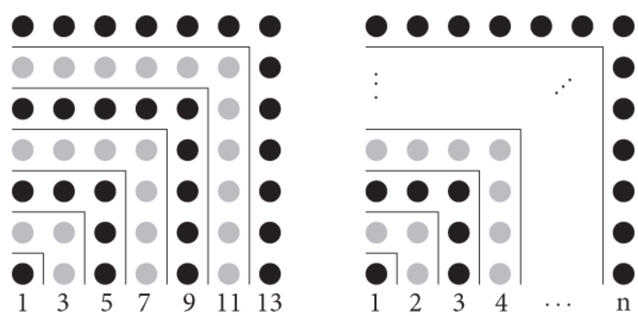
Given this, it is easy to show that the sum of digits method always works.

We demonstrated that this works for 3 previously.

Odd number theorem

Here is a simple but very useful inductive proof.

The *odd number theorem* says that the sum of the first n odd numbers is equal to n^2. Here is a "proof without words".



We prove this by induction.

$$(0 \cdot 2 + 1) + (1 \cdot 2 + 1) + \ldots \cdots + ((n-1) \cdot 2 + 1) = n^2$$

Notice that the nth odd number is $2 \cdot (n-1) + 1$.

Our formula says that

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2$$

If you like the summation style:

$$\sum_{k=0}^{n} 2k - 1 = n^2$$

As an example, the first five odd numbers are

$$1 + 3 + 5 + 7 + 9 = 25 = 5^2$$

So, if we consider the next odd number, n changes to $n + 1$. The left-hand side gets another term: we add $2 \cdot (n + 1) - 1$ to it. That is equal to $2n + 1$.

To maintain the equality, add the same quantity to the right-hand side:

$$n^2 + 2n + 1 = (n + 1)^2$$

Rearrange the result, and that's our formula back again. We have proved the inductive step.

To finish, note that the base case is simply

$$1 = 1^2$$

$\square$

The binomial theorem gives the cofactors for a binomial expansion like:

$$(a + b)^5 = a^5 + 5ab^4 + 10a^2b3 + 10a^3b^2 + 5a^4b + b^5$$

We will prove this theorem using induction **here**.

proof of induction

According to Hamming, if you are not convinced by the ladder analogy, here is another proof that induction works:

> Suppose the statement is not true for every positive (non-negative) integer. Then there are some false cases. Consider the set for which the statement is false. *If* this is a non-empty set, then it would have a least integer, which is m. Now consider the preceeding case, which is $m - 1$. This $(m - 1)$th case must be true by definition, and we know that there is such a case because as a basis for the induction we showed that there was at least one true case. We now apply the step forward, starting from this true case $m - 1$, and conclude that the next case, case m, must be true. But we assumed that it was *false*! A contradiction.

Therefore, there are no false cases.

□

Chapter 9

Sum of integers

In calculus, we will compute Riemann sums, and to do that we need to find formulas for the sum of squared integers, cubed integers, and so on. To keep it simple, let's start with the integers from 1 to n.

In a previous chapter we introduced the method called induction. Probably the most famous example of an inductive proof is that for the sum of integers.

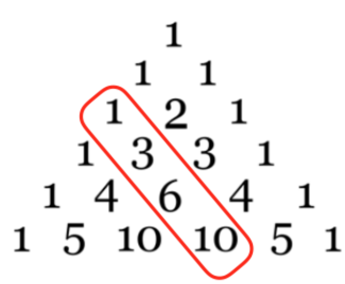
$$S_n = 1 + 2 + \cdots + n$$

proof

The numbers we seek are called the triangular numbers. These are

$$1, 3, 6, 10 \cdots$$

These are generated as the third diagonal of Pascal's triangle.:



Suppose someone has sent us, anonymously, a formula which they claim gives the sum of the first n integers, namely

$$S_n = \frac{n(n+1)}{2}$$

Assume the formula is correct for S_n. Add $n+1$ to both sides. The left-hand side becomes S_{n+1}, so we have:

$$S_{n+1} = \frac{(n)(n+1)}{2} + (n+1)$$

Rearranging:

$$= \frac{n(n+1) + 2(n+1)}{2}$$
$$= \frac{(n+1)(n+2)}{2}$$

which is exactly what we'd get by substituting $n+1$ for n in the original formula.

Alternatively, sometimes it's clearer to assume the $n-1$ case and prove the formula is correct for n:

$$S_{n-1} = \frac{n(n-1)}{2}$$
$$S_n = \frac{n(n-1)}{2} + n$$
$$= \frac{n(n-1) + 2n}{2}$$
$$= \frac{n(n-1+2)}{2} = \frac{n(n+1)}{2}$$

So we have proven that if the S_n formula is correct, then so is the one for S_{n+1}.

How do we know that S_n is correct?

Just check the *base case*:

$$S_1 = \frac{1(1+1)}{2} = 1$$

Since S_1 is clearly correct, S_2 must be also, and this continues all the way to S_n.

$$S_1 \Rightarrow S_2 \Rightarrow \ldots S_{n-1} \Rightarrow S_n \Rightarrow S_{n+1}$$

Therefore, it must be true for *every* integer n.

□

There is a famous story about Gauss. As a schoolboy, he "saw" how to add the integers from 1 to 100 as two parallel sums.

Added together horizontally, these two series must equal twice the sum of 1 to 100.

But vertically, we notice that each sum is equal to $n + 1$, and we have n of them.

$$
\begin{array}{cccccc|l}
1 & 2 & \ldots & 99 & 100 & & S_n \\
100 & 99 & \ldots & 2 & 1 & & S_n \\
\hline
101 & 101 & & 101 & 101 & &
\end{array}
$$

So, again

$$2S_n = n(n+1)$$

$$S_n = \frac{1}{2}\, n(n+1)$$

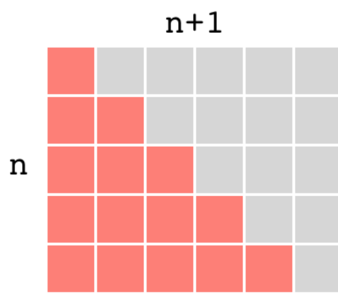
For $n = 100$ the value of the sum is 5050, which is what Gauss wrote on his slate and presented to the teacher immediately on being given the problem as a make-work exercise.

One way of looking at this result is that between 1 and 100 there are 100 representatives of the "average" value in the sequence, which (because of the monotonic steps) is $(100 + 1)/2 = 50.5$.

Or alternatively, view the sum as ranging from 0 to 100 (with the same answer). Now there are 101 examples of the average value $(100 + 0)/2 = 50$).

proof without words

Here is a striking *visual proof* of the formula to obtain T_n, the n^{th} such number. The total number of circles in the figure below is $n \times (n+1)$ and this is exactly two times the sum of the integers from 1 to n.



$$2S = n(n+1)$$

Derivation using sums

It seems a shame to spoil such a beautiful proof "without words" as the one above by saying anything more, but I can't resist. I'd like to derive the equation we have been using using algebra. The general method will help us later.

For any number, and in particular, any integer k it is true that

$$(k+1)^2 = k^2 + 2k + 1$$

So consider what happens if we sum the values from $k = 1 \to n$ for each of these terms

$$\sum_{k=1}^{n}(k+1)^2 = \sum_{k=1}^{n} k^2 + \sum_{k=1}^{n} 2k + \sum_{k=1}^{n} 1$$

If the equation is valid for any individual k, then the sum is also valid, plugging in all k up to n.

Rearranging

$$\sum_{k=1}^{n}(k+1)^2 - \sum_{k=1}^{n} k^2 = \sum_{k=1}^{n} 2k + \sum_{k=1}^{n} 1$$

Now think about the left-hand side in our equation.

$$\sum_{k=1}^{n}(k+1)^2 - \sum_{k=1}^{n} k^2$$

We have a bunch of terms starting with 2^2:

$$2^2 + 3^2 + \cdots + n^2 + (n+1)^2$$

we also have a bunch of terms to subtract starting with 1^2:

$$1^2 + 2^2 + 3^2 + \cdots + n^2$$

Almost everything cancels. This is called a "collapsing" or "telescoping" sum. We have

$$(n+1)^2 - 1 = n^2 + 2n$$

Bringing back the right-hand side we obtain:

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

We can bring the constant factor 2 out of the sum, and also, we recognize that the sum of the value 1 a total of n times is just n.

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

Subtract n from both sides and divide by 2:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

That's it!

Part III

Lines and triangles

Chapter 10

Lines and angles

Euclid and the postulates

Greek geometry starts hundreds of years before Euclid, who was a contemporary of Alexander the Great (356-323 BC).

We know that Euclid lived after Plato (died 347 BC), and before Archimedes (born 287 BC). Except that he worked in Alexandria, all other details of his life and death are shrouded in mystery.

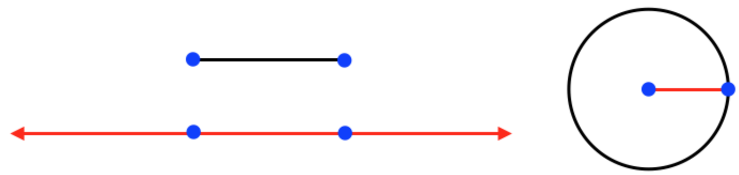
After more than 2000 years, Euclid's book *Elements* is still an excellent place to begin surveying the foundations of geometry. It is a textbook, an organized collection of everything that a well-educated student was expected to know about the subject at the time.

This book of consists of *propositions*, which include constructions (geometric figures) drawn with a pencil on a piece of paper, using a straight-edge or a compass or both.

Often proofs of propositions build on previous items in the book. Euclid does not prove everything. Bertrand Russell was famously disappointed about that.

Here are Euclid's first three postulates — statements that he assumed to be true:

• A straight line segment can be drawn joining any two points.



• Any straight line segment can be extended indefinitely in a straight line.

• Given any straight line segment, a circle can be drawn having the segment as the radius and one endpoint as the center.

Let us assume these as well. We will use them often.

We finesse the difficulty in defining what is meant by *straight* in the real world. If you've ever done any carpentry, you probably know that unknown edges are determined to be straight by comparison with another edge known to be straight.

In geometry, we use an imaginary perfect straight-edge to draw a straight line as "the shortest distance between two points".

The fourth postulate is:

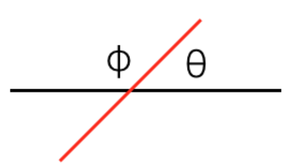
• All right angles are congruent, that is, equal to each other.

This one prompts two different questions. The first one is:

◦ how do we measure an angle?

measure of an angle

Consider the diagram below. One line segment is drawn crossing a second one, forming their *intersection*



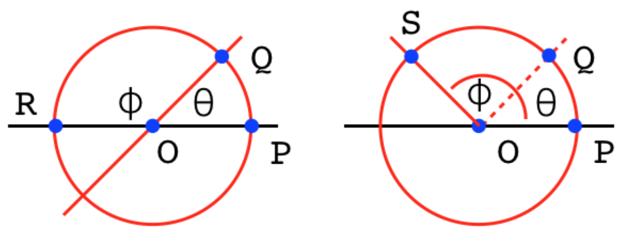
Two angles are labeled. One, ϕ, is larger than the other one, θ.

$$\phi > \theta$$

Writing the statement $\phi > \theta$ is easy, but the implication is that we have some way of talking about the *measure* of an angle.

Our answer is to construct a circle around the central point, and call the distance along the circumference between the points where the lines cross the circle, the measure of the angle.

If that distance along the edge is larger for $\angle\phi$, then $\phi > \theta$. In the left panel, the arc between Q and R (call it arc QR) is larger than arc PQ.



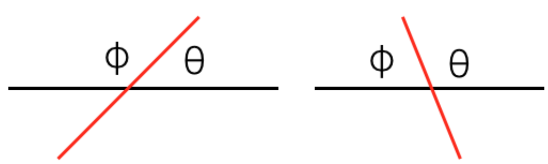
We don't need to actually measure the arc.

Instead we can use a standard compass to lay off the linear distance from Q to R starting at P (right panel). Since S is further around the circle, $\phi > \theta$.

As for our second question:

- what is a *right angle*?

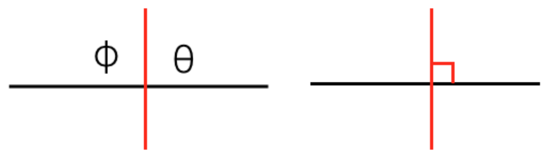
Let us call angles formed on the same side of a line, *supplementary* angles (also sometimes called adjacent angles).



On the left, one of the angles, ϕ, is larger than the other one, θ. On the right, we have $\phi < \theta$.

There is of course a third possibility, namely that $\theta = \phi$. The definition of a right angle is that

- if two supplementary angles are equal, then they are both right angles.



A right angle is frequently designated by drawing a small square, as seen in the right panel above.

Regardless of whether $\theta < \phi$, $\theta > \phi$, or $\theta = \phi$, the sum of the two angles $\phi + \theta$ is equal to two right angles or 180 degrees.

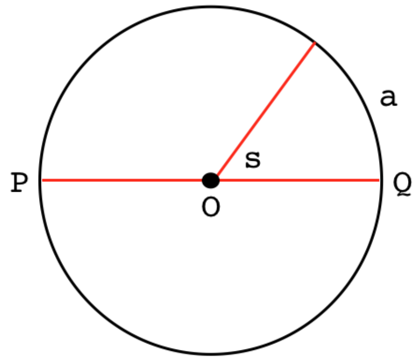
The convention that there are 360 degrees total in a circle dates to the time of the Babylonians (c. 2400 BC).

There is nothing particularly special about using 90 degrees as the measure of a right angle, 180 degrees for any two supplementary angles including right angles, or 360 degrees for one whole turn.

Well, there is one thing: there are *approximately* 360 days in a year, which marks the sun's track across the sky.

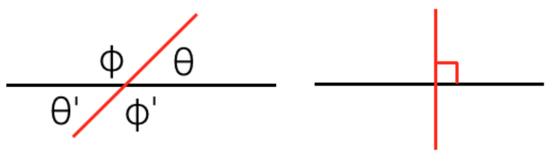
In his book, *Measurement*, Lockhart adopts the convention that a whole turn is equal to 1.

We'll just mention here that one whole turn can be defined using a different unit of measure as 2π *radians*, and that convention turns out to be quite important for calculus.



In the figure above, we *define* the measure of the angle s to be equal to the arc it sweeps out or subtends, a, in a circle of radius 1, a *unit* circle. Angles are not lengths, but numerically, the measure of the angle is the measure of the arc.

Now, consider those angles lying below the horizontal:



We said that the sum of the two angles $\phi + \theta$ is equal to two right angles, but so are the sums $\theta' + \phi$ and $\theta + \phi'$, for the same reason. As a result

$$\phi + \theta = \theta + \phi'$$

We conclude that

$$\phi = \phi'$$

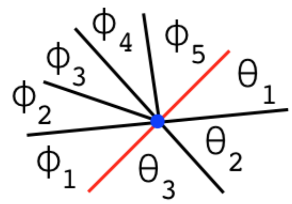
and

$$\theta = \theta'$$

This is called the *vertical angle theorem*.

On the right, if any one of the angles where two lines cross is a right angle, then all four are right angles.

Finally, the idea is generalizable to more than just two angles. In the figure below



$$s = \phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5$$

$$t = \theta_1 + \theta_2 + \theta_3$$

$$s = t = \pi$$

$$s + t = 2\pi$$

Both sums are equal to two right angles.

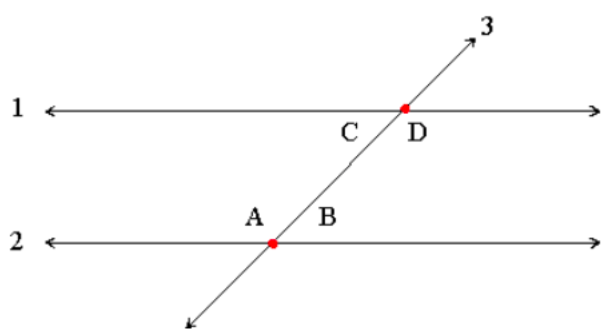
- when one or more lines cross a given line at the same point, the angles formed on one side of the given line have a total measure equal to two right angles.
- the total of all the angles formed at that point is equal to four right angles.

parallel postulate

So far, all this seems rather obvious. The fifth and final postulate is more subtle.

In the figure, line 1 and line 2 are parallel, *if and only if*

$$A + C = B + D = 180 = 2 \text{ right angles}$$



• If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough.

This postulate is equivalent to what is known as the parallel postulate.

http://mathworld.wolfram.com/EuclidsPostulates.html

alternate interior angles

We come to a very important theorem.

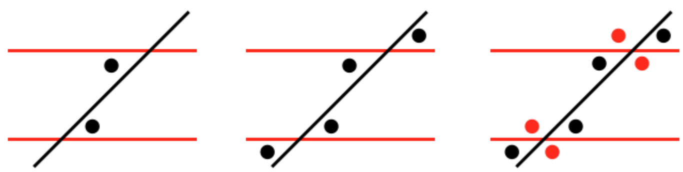
In the figure above, the supplementary angles $A + B = 180$ also add up to 180 degrees. So

$$A + B = 180 = A + C$$

and then

$$B = C$$

This is called the theorem on *alternate interior angles* between two parallel lines.



In the figure above (left panel), we're given that the two horizontal lines are parallel.

The indicated angles are equal because they are alternate interior angles of two parallel lines (parallel postulate).

In the middle panel, two additional equalities are established by the vertical angle theorem. Then on the right, we use the supplementary angle theorem.

Note that the conclusions for the angles marked with a red dot are also themselves consistent vertical angles and alternate interior angles.

summary

Make sure you learn and understand each of these theorems:

• supplementary angles

- vertical angles
- alternate interior angles
- triangle sum

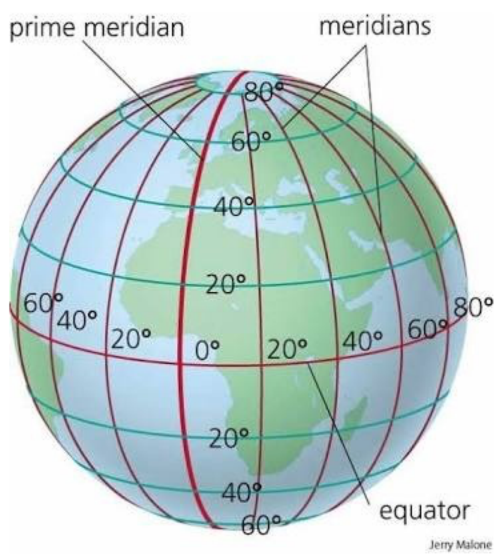
Another point to remember: these are two-way, *if and only if* theorems. So for example, if the two alternate interior angles of a traversal are equal, the two lines are parallel.

flat geometry

Adoption of the parallel postulate is a choice. The definition works for geometry in the flat plane.

But consider a familiar situation where this is not true. Suppose we are doing geometry on the surface of a sphere, such as the earth.

Then, two adjacent lines of longitude can be drawn so as to cross the equator at right angles, and the lines are parallel there, but they will meet (intersect) at the poles.



The parallel postulate only holds for geometry on a *flat* surface.

axioms

Euclid also lists five axioms, things which are assumed. Here are two examples:

• Things that are equal to the same thing are also equal to one another.

• If equals are added to equals, then the wholes are equal.

These seem quite reasonable.

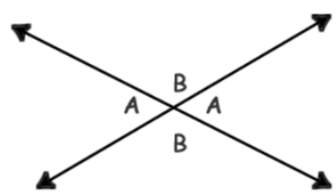
We will see how to proceed from the postulates and axioms to various proofs. Given these *assumptions*, we can prove theorems that must be true.

Thales

I'm a big fan of William Dunham's books — several of them are listed in the References.

Dunham has written a lot about the history of mathematics in Greece, starting with Thales (624-546 BC), who was from a Greek town called Miletus on the coast of Asia Minor (modern Turkey). He lived long before Euclid (about 300 years before, 600 BC). Although none of his writing survives, it is believed that Thales proved several early theorems including one we saw above.

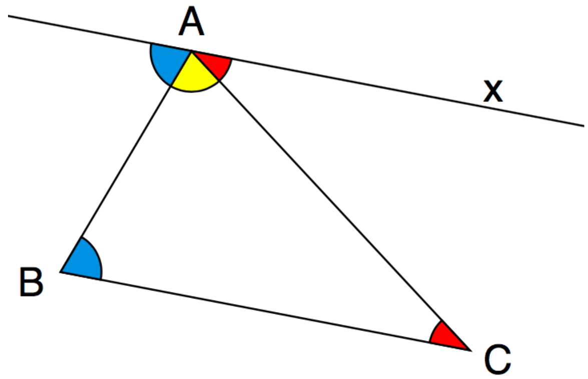
• The vertical angles formed by two straight lines crossing, are equal.



This theorem, which we already proved, depends on a property of straight lines. In the proof, we used the axiom "equals added to equals are equal", alternatively "equals subtracted from equals are equal."

A very important theorem attributed to Thales is the following:

• The angle sum of a triangle is equal to two right angles.



This theorem depends on the ideas we developed above. Draw a line segment through A parallel to BC. Now, use alternate interior angles and follow the colors to the result.

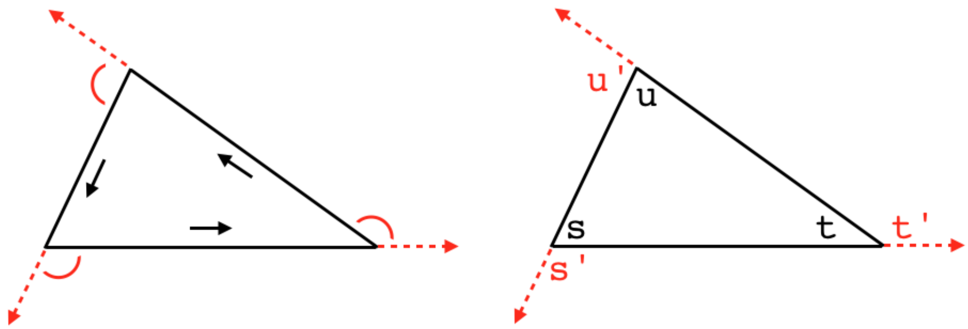
We will see one more theorem ascribed to Thales (it is actually called Thales' theorem), in the next chapter. It is about isosceles triangles (two sides equal).

another proof

Here is a different proof of the theorem on the sum of angles in a triangle adding to 180 degrees. It never hurts to re-prove things by a different method. It serves as a check on both the result, and the methods.

Imagine walking around the perimeter of a triangle in the counter-clockwise direction. At each vertex we turn left by a certain number of degrees, θ, called the exterior angle. After passing through all three vertices, we must end up facing in the same direction as we started.

The sum of the exterior angles is 360°.



$$s' + t' + u' = 360$$

In addition, for each vertex, the interior angle plus the exterior angle add up to 180 degrees. If we add all three pairs, we obtain

$$(s + s') + (t + t') + (u + u') = 3 \cdot 180 = 540$$

By subtraction

$$s + t + u = 180$$

Chapter 11

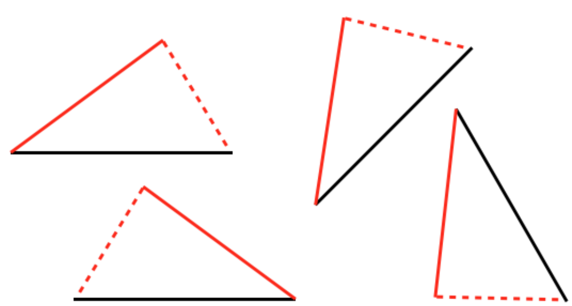
Congruent triangles

congruence

• Two triangles are *congruent* if and only if they have the same three side lengths.

Congruence is a fancy word for equal. The condition we use is often abbreviated SSS (side-side-side).

By this definition, a triangle and its mirror image are congruent. The three triangles shown below are all congruent, even though two are flipped (they are mirror images).



Having three sides equal means that the shape is the same. The three angles are also equal — and the shapes are superimposable, with the proviso that we allow the shape to be flipped over.

In addition to SSS (side-side-side), there are three other conditions that lead to congruence of two triangles when they are satisfied, namely

• SAS (side-angle-side)

• ASA (angle-side-angle)

• AAS (angle-angle-side)

similarity

Some triangles are *similar* but not congruent.

Similarity means that the three angles are the same but the triangles are of different overall sizes. We might say that they are the same but *scaled* differently.

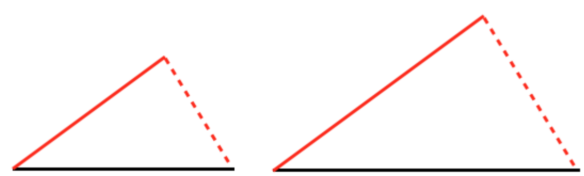
We can call this AAA (angle-angle-angle). For similar triangles, the three corresponding pairs of sides are in the same proportions, but re-scaled by a constant of proportion.

• Two triangles are similar if they have the same three angles.

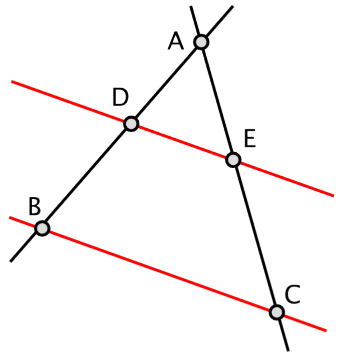
Because of the angle sum theorem, if any two angles of a pair of triangles are known to be equal, then the third one must be equal as well.

• Two triangles are similar if they have two angles known to be equal.

Similar triangles have their sides in the same proportion. This is known as the AAA similarity theorem.



Given any triangle, draw a line parallel to one side, which also joins the other two sides. The new triangle with that side as its base is similar to the given triangle.



In this example, these ratios are equal

$$\frac{AD}{AB} = \frac{AE}{AC} = \frac{DE}{BC}$$

and you can find others.

Statements about similarity:

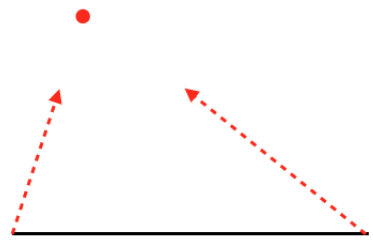
- similar triangles have all three angles equal
- if two similar triangles are superimposed, the two sides that do not coincide with each other, are parallel
- similar triangles have their similar sides in the same ratio

It is easy to see why the first two statements are equivalent. Just use the alternate interior angles theorem.

The third is harder. For now we will assume the theorem: that AAA and sides in proportion are both the same as similarity.

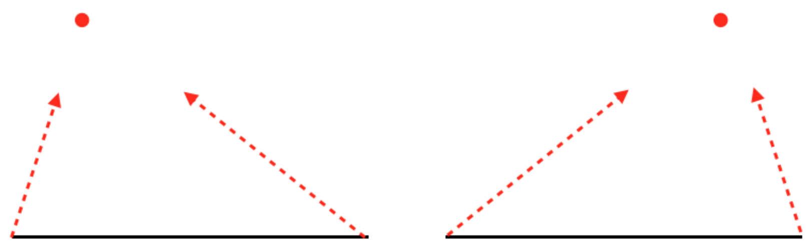
constructions

The way I think about the congruence conditions is to imagine trying to construct a triangle from the given information, and ask whether it is uniquely determined. Suppose we know ASA. The situation is thus:



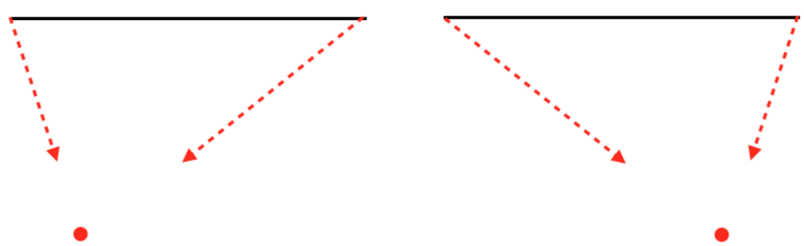
Draw the known side, then using the known angles, start two other sides from the ends of that side. They must cross at a unique point.

But... actually, if we start the two lines from opposite ends of the horizontal



there is another solution, the mirror image. These two triangles are congruent to the one above.

I'm tempted to add still another:



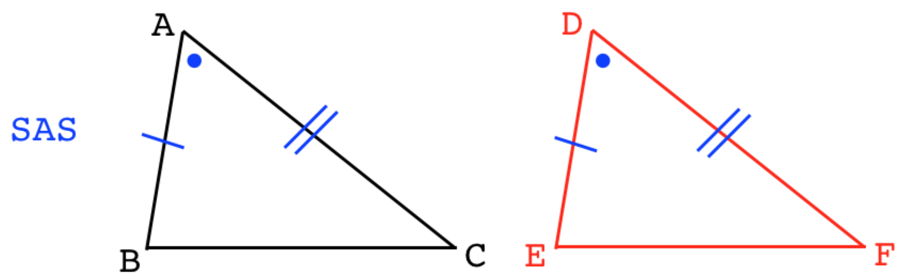
But this doesn't give anything new. These are merely rotated versions of the ones above. Congruent triangles include the two mirror images and that's it.

If we know two angles we also know the third, because they must add to 180 degrees. For this reason, ASA and AAS imply that we have exactly the same information, because we know all three angles and we know one side.

This restriction is crucial: we must also know *which* two angles flank the known side. Alternatively, it is enough to know which angle is opposite to the known side.

SAS, ASA, AAS but not ASS

SAS is very commonly used to prove congruence.

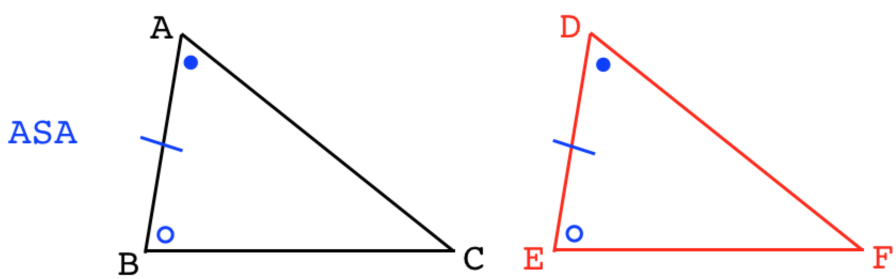


In this diagram, sides of equal length are indicated by one or more hash marks.

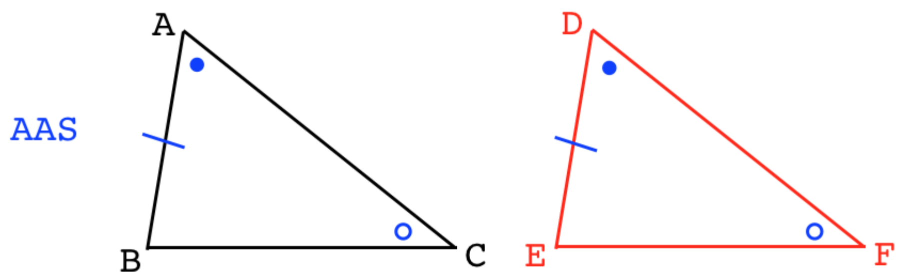
Equal angles are usually indicated by dots in this book. (Dots are easier to place on the figures, and lend themselves to color-coding; the common method for pencil and paper is to draw an arc with a hash across it).

The other methods for proving congruence use two equal angles and a side. Two equal angles imply the third angle is also equal (since they add to a half-circle or 180 degrees), so the two triangles are similar. To prove they are congruent, we need one side.

These methods using two angles are referred to as ASA



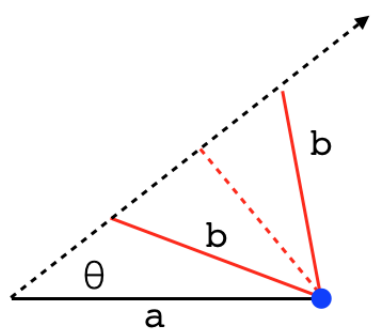
and AAS.



Because AA implies AAA, these tell unambiguously which angle is opposite the known side.

one that doesn't work

There is one set of three that doesn't work in the general case, that is ASS (angle-side-side).



Here we know sides a and b and the angle θ adjacent to a and facing opposite side b. Imagine b swinging on a hinge at the blue dot. If $b < a$, there are two points where b can intersect with the side projecting from angle θ. There is no unique solution, so the triangle is not determined.

If it had been the case that $b > a$, or alternatively that b formed a right angle with the third side, then the triangle *would* be determined.

I think Tony Randall said it best

https://www.youtube.com/watch?v=KEP1acj29-Y

Chapter 12

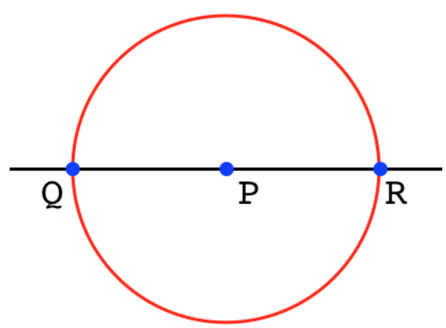
Perpendicular bisector

perpendicular at a point

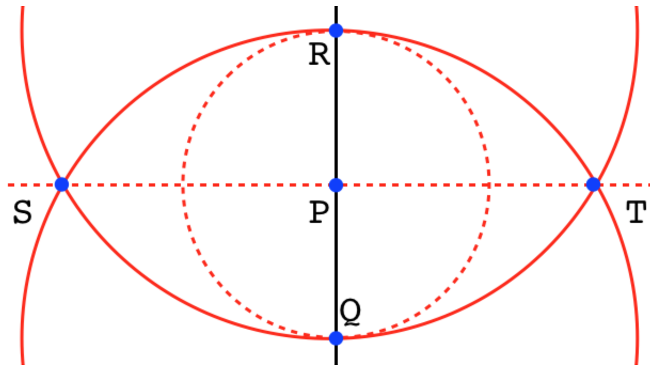
It's common to want to construct a line segment perpendicular to another line segment. The perpendicular might be specified to occur at a particular given point, either on the line, or alternatively through some other point not on the line.

For the first case, consider the horizontal line segment below. Suppose we know a point P on the line and wish to construct the vertical line through P.

Use the compass to mark off points Q and R on both sides, an equal distance from P. This can be done by drawing a circle with center at P. The radius is PR (also equal to QP).



Find S equally distant from Q and R. This can be done by using the compass to draw larger circles of the same radius on centers Q and R. Since we have the room, I've drawn much larger circles of radius equal to QR.



The figure is rotated by 90 degrees to conserve vertical space on the page.

$QS = RS$ because they are both radii of circles of equal radius.

The line segment SP will be perpendicular to the line containing QPR, because of theorems about triangles with two equal sides (isosceles) which we will prove later.

collapsible compass

Note briefly that is a restriction in Euclid's *Elements* to a *collapsible* compass, which is a compass that loses its setting when lifted from the page. That means that generally, you wouldn't be able to draw two circles of the same radius on different centers.

We got around that restriction by drawing the circles on Q and R with the same radius QR.

We will call a compass that is able to hold its setting, a *standard* compass, and explain why the distinction is important to Euclid in the chapter devoted to his book. But we also note that within the first few pages of that book, it is shown how to use a collapsible compass to carry out the very construction we said we couldn't do, namely, construct two circles on Q and R with equal radius and that radius not equal to QR or QP.

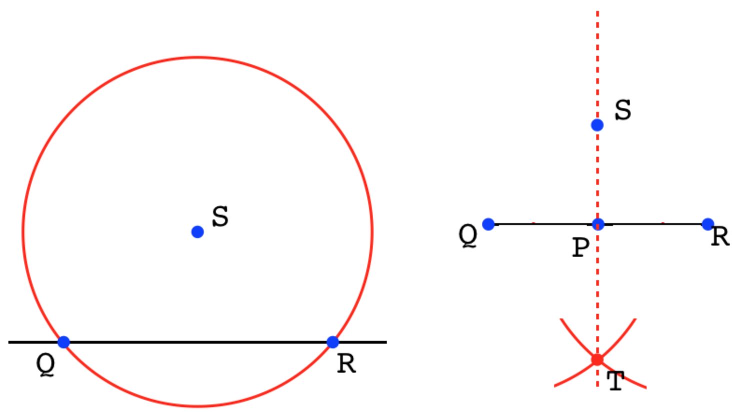
bisect a line segment

Suppose that we had not known the point P when we started the procedure above, but already had two points Q and R.

Then the line through S and T crosses QR as a perpendicular at its midpoint, and we have found the point P that bisects QR.

perpendicular through a point

Alternatively, suppose we know the line and the point S but not P, and we wish to construct a vertical through the line that also passes through S. Find Q and R on the line an equal distance from S ($QS = RS$), as radii of a circle centered at S (left panel, below). Their exact position is unimportant.



Now repeat the previous construction, using Q and R.

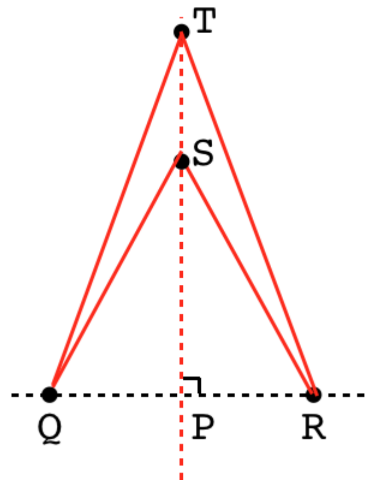
The line segment ST is perpendicular to the line segment containing QR, and passes through S, as required.

Also, see the video at the url:

https://www.mathopenref.com/constperpextpoint.html

bisector properties

Using what we've just learned, suppose we know two points Q and R. We find the point P equidistant between them and construct the perpendicular bisector PS. Then the two sides SQ and SR have equal length. Triangle $\triangle SQR$ is isosceles.



Proof.

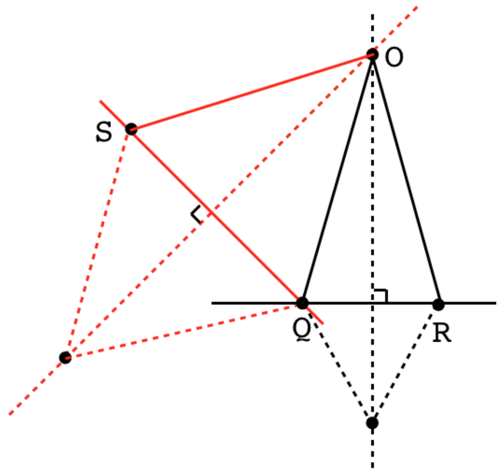
By construction, $PQ = PR$, $\angle SPR$ is a right angle, and side SP is shared. Hence the two triangles $\triangle SPQ$ and $\triangle SPR$ are congruent, by SAS.

$\square$

This is true for *any* point on the line drawn through S and P. For example, $TQ = TR$ in the figure above.

three points

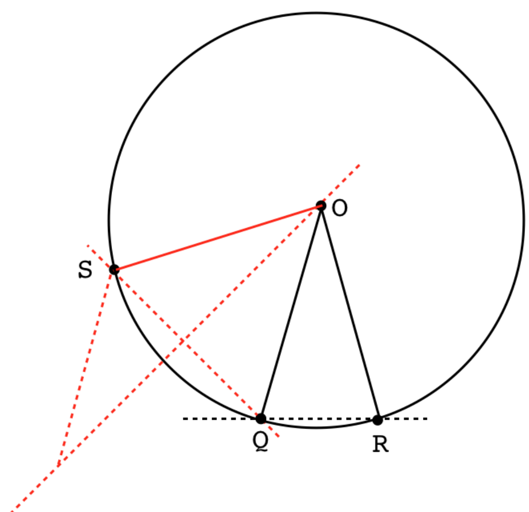
Now, suppose we have three points: Q, R and S. We find the perpendicular bisector of QR and also, the perpendicular bisector of QS. Extend them to where they meet, at point O.



What can we say about point O?

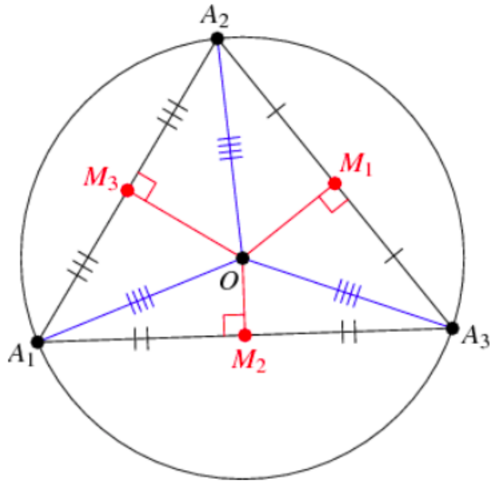
- O is equidistant from Q and R.
- O is also equidistant from Q and S.

Therefore, $OQ = OR = OS$. If we draw a circle on center O with radius OR, it will pass through all three points.



circumcenter

The point where the perpendicular bisectors cross has a special name, it is called the circumcenter.

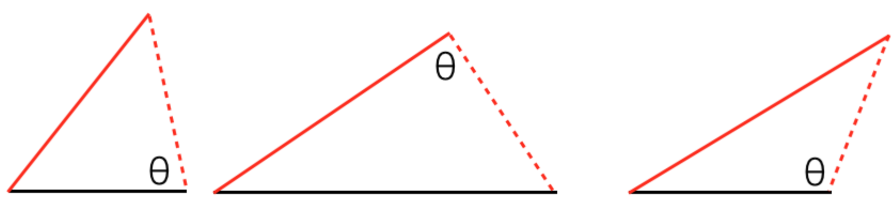


There are other special points where interesting circles can be drawn. We'll talk about them in a bit.

Chapter 13

Special triangles

There are several adjectives one can use to describe different types of triangles. For example: acute, right, and obtuse.



The acute triangle (left) has all three angles smaller than a right angle. The right triangle, naturally, has one right angle.

We'll say a lot more about right triangles later.

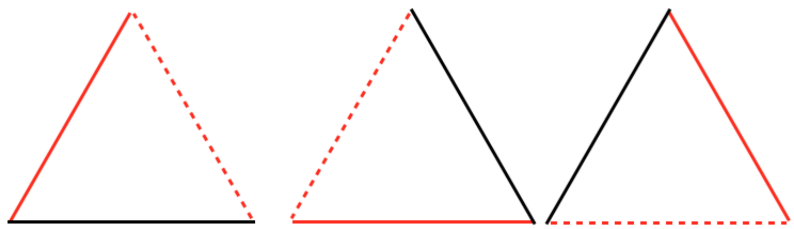
Finally, an obtuse triangle has one angle larger than a right angle (right, above).

symmetry

One can also talk about the situation where either two sides, or all three sides, have the same length.

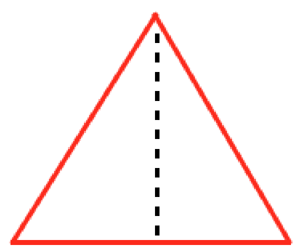
An equilateral triangle has all three sides the same, while an isosceles triangle has two sides the same length.

The most important consequence of three sides equal for an equilateral triangle, is rotational symmetry. Three turns of 120 degrees, and we're back where we started.



The implication of that is that the three angles are also equal. There is no reason to choose one larger than any other.

In the next figure the two smaller triangles obtained by dividing in half an equilateral triangle (all sides equal), are congruent.



By divide in half, we mean bisect the base and draw the line from the top vertex. We have SSS.

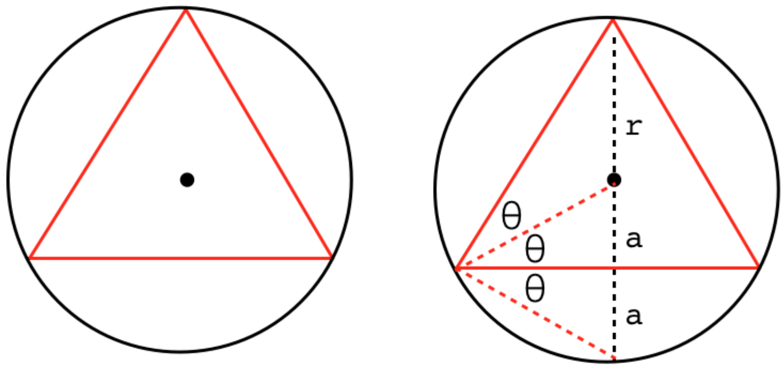
We also have that the two angles at the base of the bisector are equal and supplementary angles. Therefore, they are both right angles.

□

circumscribed circle

Here is a fun construction based on an equilateral triangle.

Any triangle fits into a unique circle. We will prove this elsewhere.



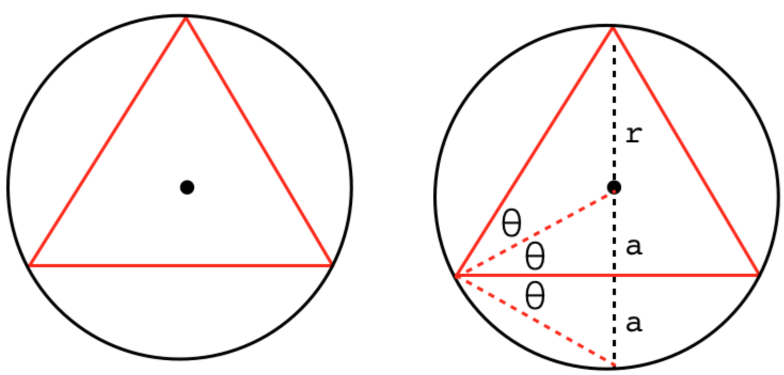
If we draw the radius to a vertex of the triangle, and then to the end of the diameter, it makes you wonder whether the three angles marked θ are equal.

Proof.

First, the radius is an altitude, and it divides the vertex angle in half, as we've been saying. So that accounts for two of the angles labeled θ.

The third comes about because any triangle with two points at the ends of the diagonal and a third anywhere on the circle is a right triangle. We'll prove that when we get to circles.

As a result, θ measures 30 degrees. Since a right triangle is 90, we assign the third θ.



So now we have a smaller angle of a right triangle, and a shared side. The two triangles are congruent. That accounts for the duplicated a in the figure.

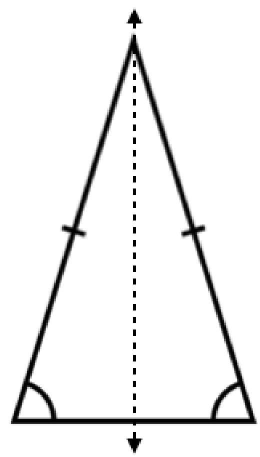
Thus, the altitude of the equilateral triangle is 3/4 of the diameter of the circle that just encloses it. And the point where the altitudes meet in an equilateral triangle is

1/3 of the way up from the base, since $r = 2a$.

We will see that such a point where the altitudes cross is unique and exists for any triangle, and it always has the same measure as a fraction of the altitude. This is called Ceva's theorem.

theorem from Thales

• The base angles of an isosceles triangle are equal. Also, if the two base angles are equal, the triangle is isosceles.



My favorite proof of this theorem is from reflective or mirror image symmetry (above).

Start with the two sides equal and draw a line to the midpoint of the base opposite. The figure has reflective symmetry, thus the angle is bisected.

We prove this more carefully now.

notation

The Greeks, including Euclid, adhere to certain conventions. For example, points are always labeled with letters, line segments are referred to by the endpoints, and angles by the line segments that determine them, as in $\angle ABC = \angle DEF$.

I don't know about you but I find myself tracing out angles from the three points, again and again.

We could give labels to the angles like $\alpha, \beta \ldots$ and so on, to the sides opposite vertices as a opposite A and so on.

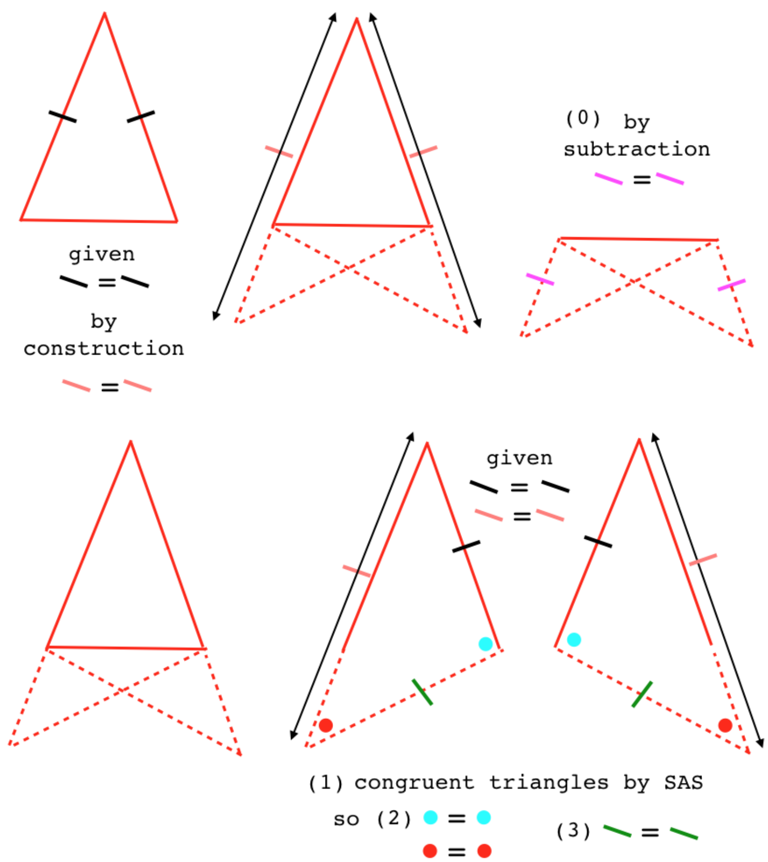
Boldly, we choose to be even more dramatic. Dispense with labels altogether and use colored dots for equal angles and colored bars for equal lengths.

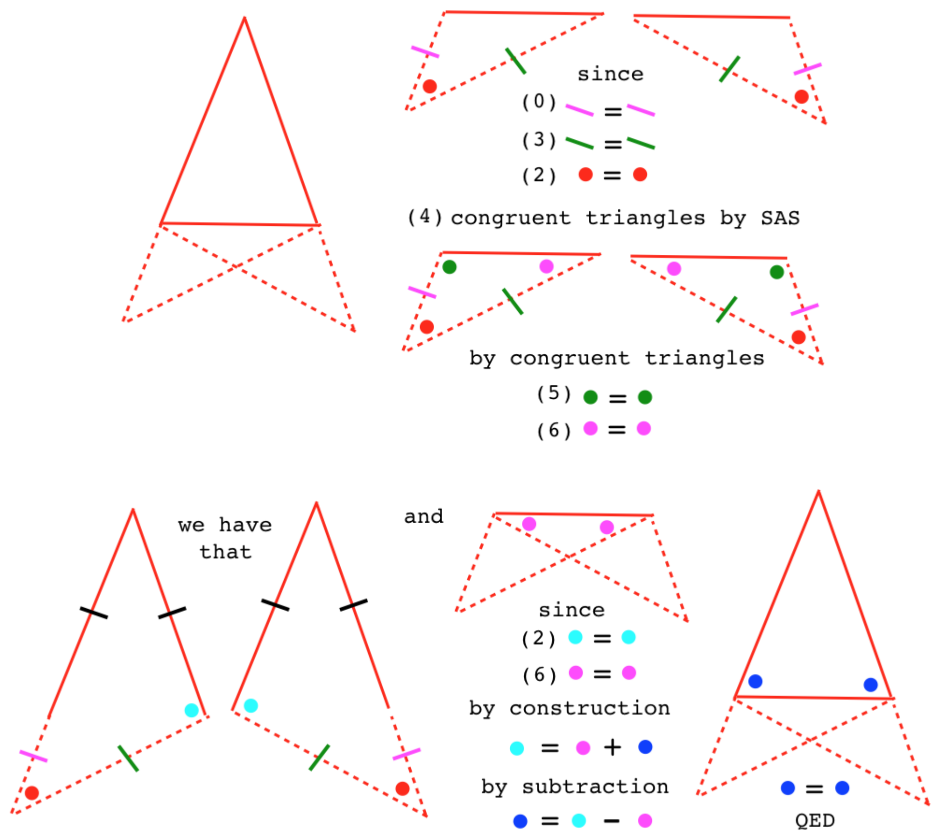
Here is the famous proof of Thales' theorem from Euclid's *Elements*.

Prop. I.5

In isosceles triangles the angles at the base equal one another, and, if the equal straight lines are produced further, then the angles under the base equal one another.

In what follows, all the pieces are with reference to the initial construction, first figure, below.





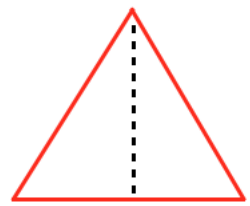
□

The theorem says that the base angles are equal $\iff$ the two sides sides are equal (not the base).

The symbol $\iff$ means *if and only if*, so $A \iff B$ means that both $A \to B$ and $B \to A$.

So far, our proof runs only in the forward direction. We will do the converse later.

Above we said that in this figure the two smaller triangles obtained by dividing an equilateral triangle in half, are congruent. The dotted line is called an *altitude* of the triangle.



An altitude meets the side opposite in a right triangle.

Because the left and right sides of the original triangle are equal, the base angles are equal, by the property of isosceles triangles which we just proved. The angles where the altitude meets the base are both right angles, by symmetry and by the definition of the altitude.

Therefore we have AAS, and the two halves are congruent.

So, the two angles at the top where the altitude meets the sides are also equal (as the third angle with the other two angles determined).

Chapter 14

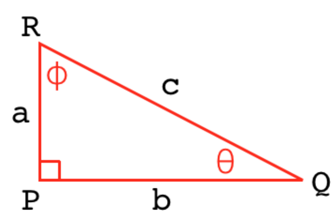
Right triangles

The main result we are headed for is the Pythagorean Theorem. Before we get there, however, it is worthwhile to continue our development of basic geometry with a discussion about right angles and right triangles.

A right triangle is a triangle containing one right angle. Right angles (and right triangles) are special. We saw previously that the definition of a right angle is that two of them add up to one straight line or 180 degrees.

Since we proved that the sum of the three angles in any triangle is equal to one straight line, by extension, the sum of angles in any triangle is also equal to two right angles.

In the figure below, the angle at vertex P is a right angle. It is common to mark a right angle with a little square, as shown, but these are a pain to draw, so I will often not do that. The side opposite P, namely c, is the *hypotenuse*.



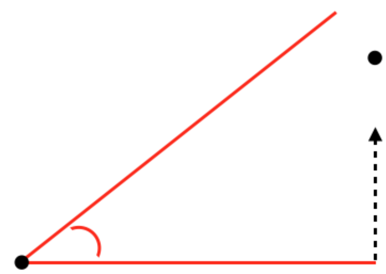
Since the sum of angles in a triangle is equal to two right angles, the sum of the angles θ and ϕ above is also equal to one right angle, or 90 degrees.

Angles θ and ϕ are said to be *complementary*. This fact is often exploited in proofs.

- the two smaller angles in a right triangle are complementary and add to 90 degrees.

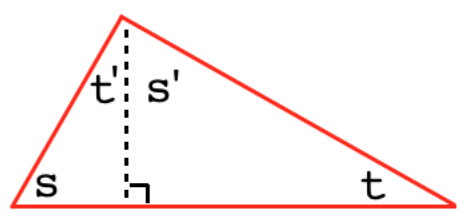
right triangles

For two right-triangles, if one hypotenuse is equal to the other, and also one set of legs equal, the two triangles are congruent.



In the figure, imagine the hypotenuse swinging on the hinge of its vertex with the horizontal base. There is only one angle where it will terminate on the vertical side with the correct length. This determines the angle between the known sides, or alternatively, the length of the third side.

altitudes



In the large right triangle above, we know that

$$s + t = 90$$

When we draw the perpendicular to the hypotenuse that goes through the upper vertex, that is an *altitude* of the triangle. Because of the right angle, we obtain two smaller right triangles. Thus

$$s + t' = 90$$

$$s' + t = 90$$

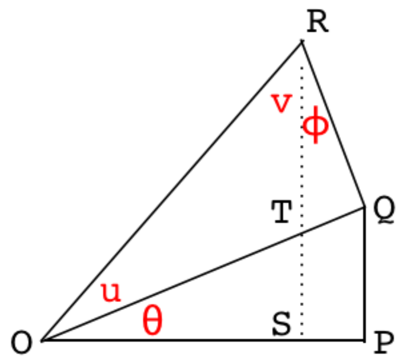
Hence

$$s + t = s + t'$$

$$t = t'$$

and similarly for s and s'.

stacked triangles



Suppose we are given that $\angle OPQ$ and $\angle OQR$ are right angles. We draw the altitude RS and observe that the angle at vertex S is a right angle.

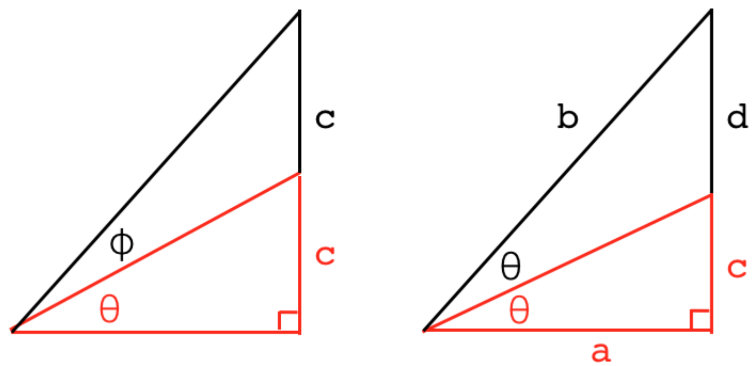
Therefore, in triangle ORS, the sum $\theta + u + v$ is equal to one right angle. At the same time, in triangle OQR, the sum $u + v + \phi$ is also equal to one right angle. Therefore

$$\theta = \phi$$

Further, $\triangle QRT$ and $\triangle OPQ$ are similar triangles.

angle bisector

With that background, we now consider a classic problem: angle bisectors.



Suppose we are given that the large triangle is a right triangle.

We draw a line joining the vertex on the left with the side opposite.

This line could in general be drawn anywhere, however two interesting cases are when the side opposite is bisected (left panel), or when the angle at the left is bisected (right panel). These two cases are not the same. In the first $\phi \neq \theta$ and in the second, $c \neq d$.

Suppose we choose the second possibility, equal angles. We are in a position to prove an important theorem.

angle bisector theorem

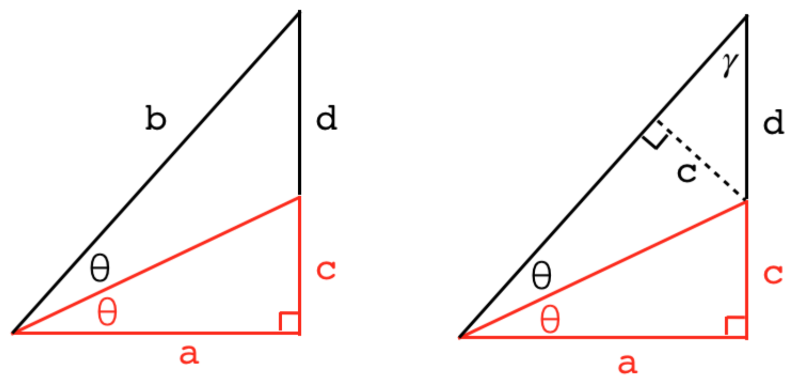
With reference to the two figures above, we are to prove that

$$\frac{d}{b} = \frac{c}{a}$$

The sides and bases are in proportion for a right triangle with bisected angle.

Proof.

Draw an altitude for the upper of the two small triangles, meeting the side of length b.



The red triangle and the one directly above it are congruent (right panel). They share a side (the hypotenuse of each), and they are right triangles with the same smaller angle γ. Therefore, the altitude we just drew has length c.

The small triangle with sides c and d (at the top) is similar to the original large triangle. The reason is that they are both right triangles containing the smaller angle γ.

By similar triangles, we form equal ratios of the angle opposite γ to the hypotenuse:

$$\frac{a}{b} = \frac{c}{d}$$

This is rearranged simply to give

$$\frac{d}{b} = \frac{c}{a}$$

which is what we were asked to prove.

□

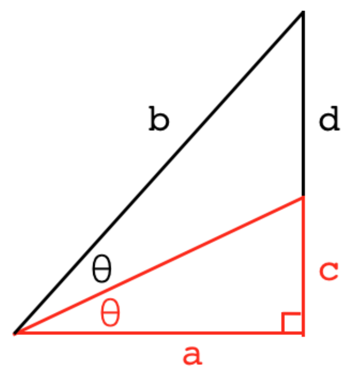
The result can be pushed a little further:

$$\frac{a}{b} = \frac{c}{d}$$

Here's the key point

$$\frac{a+b}{b} = \frac{c+d}{d}$$

$$\frac{a+b}{c+d} = \frac{b}{d} = \frac{a}{c}$$



which is a surprising result and becomes important later in looking at Archimedes method for approximating the value of π.

Chapter 15

Euclid's Elements

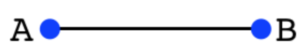
In this chapter we will study some nine or ten *Propositions* from the first volume Euclid's *Elements*. We also prove the *external angle theorem*.

The book was put together as a compendium of geometry for students. One thing we will see is how the propositions build on one another.

The first three propositions are *constructions*, e.g. the very first asks us to construct a triangle with all three sides equal, an equilateral triangle. The first statement below is Euclid's voice.

Prop. I.1

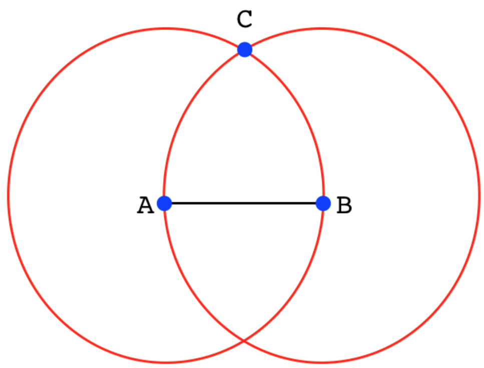
To construct an equilateral triangle on a given line segment.



The tools we have are a straight-edge and a compass. The compass is collapsible, meaning that it cannot be used to transfer distances since it loses its setting when lifted from the page. As we'll see in the next part, this is a problem with a solution.

Euclid was smart enough to know about compasses and how to set them. The idea he had was this: to make the fewest possible assumptions. A non-collapsible compass was a luxury he didn't need, since he could accomplish the same end without it, as we will see.

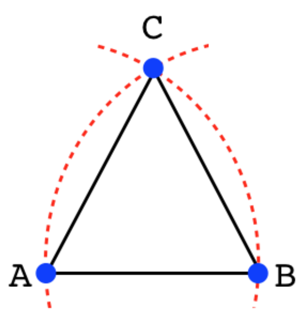
The first step is to draw two circles on centers A and B.



The circles are drawn with each radius equal to the line segment AB. It is a property of circles that all points on the circle are at the same distance from the center. Thus all points on the left-hand circle are equidistant from A, and all points on the second one are equidistant from B.

Therefore, the point C where the circles cross is equidistant from *both* A and B.

For this, we don't really need the entire circles, just the part where the arcs cross at C.



Now use the straight edge to draw $\triangle ABC$. Since $AC = AB$ and $BC = AB$, we know that $AC = BC$. The triangle is equilateral.

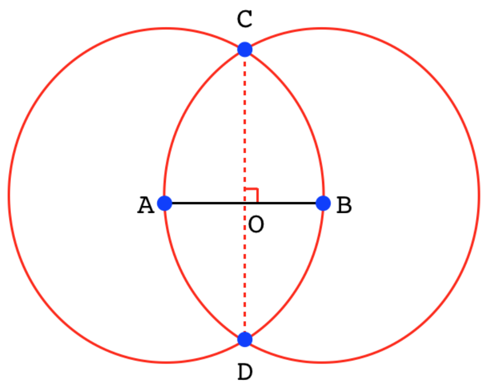
We put a little box to show that the proof is complete.

$\square$

The proof doesn't stand on its own. We used one definition (D) and a common notion (CN).

- D I.15 all radii of a circle are equal.
- CN I.1 things which equal the same thing also equal one another.

If we look again at the figure, and label the other point where the circles cross as D:



Note: CD is the perpendicular bisector of AB. Euclid doesn't have the tools to prove that yet, so he leaves it for now.

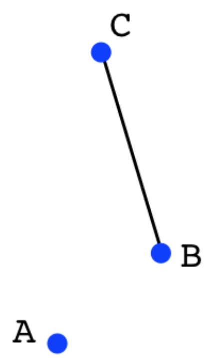
If desired, we could draw on tools from the last chapter. Clearly $\triangle ABC$ and $\triangle ABD$ are equilateral triangles with sides of the same length. From that, we deduce that $\triangle ACD$ and $\triangle CBD$ are isosceles. And from that, we can easily prove that all four small triangles with O as a vertex are equal and therefore right angles.

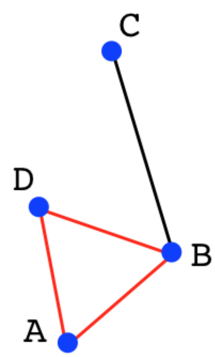
We leave this as an exercise.

Prop. I.2

To place a straight line equal to a given straight line with one end at a given point.

We will construct a line segment at A equal in length to BC (left panel). The first thing is to draw the line segment AB and construct an equilateral triangle on it (right panel).



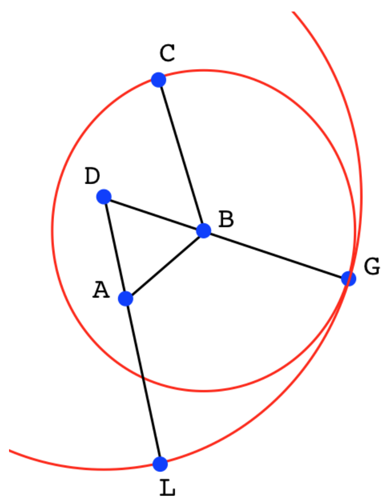


We know how to do this (from $PI.1$).

Next, construct a circle on center B with radius BC and extend the line segment DB to point G.

Then, construct a circle on center D with radius DG and extend DA to that circle at point L.

We have:



As common radii of the circle on center B, we have $BC = BG$.

As common radii of the circle on center D, we have $DL = DG$.

As sides of an equilateral triangle, we have $DA = DB$.

We use CN $I.3$: if equals are subtracted from equals, then the remainders are equal. Thus, $AL = BG$. But we had above that $BC = BG$. Therefore, $AL = BC$, by CN $I.1$.

Q.E.D. or "quod erat demonstrandum", in Latin

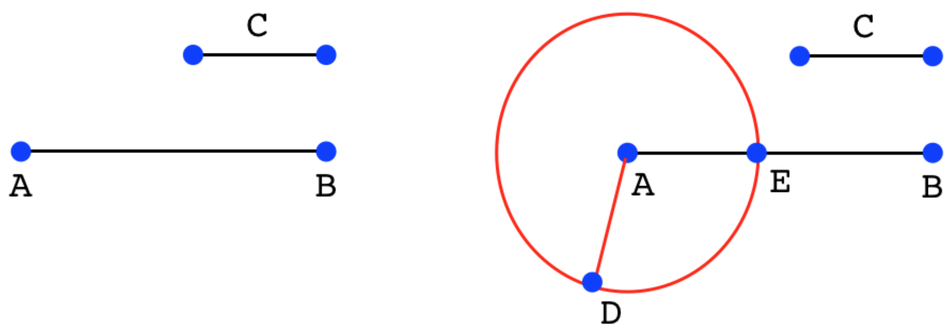
And in the original Greek *the very thing it was required to have shown.*

□

Note in passing, the orientation is determined by AB. We have not shown how to transfer the length with an arbitrary orientation. We will solve this next.

Prop. I.3

To cut off the lesser of two unequal straight lines from the greater.



In the left panel, we have the line segment AB and a smaller one just labeled C. To do the construction, use the method of P $I.2$ and transfer C to point A, forming AD.

Next, use AD as the radius of a circle on center A. Then, $AE = AD$, but $AD = C$. Hence

$$BE = AB - AE = AB - AD = AB - C$$

as required.

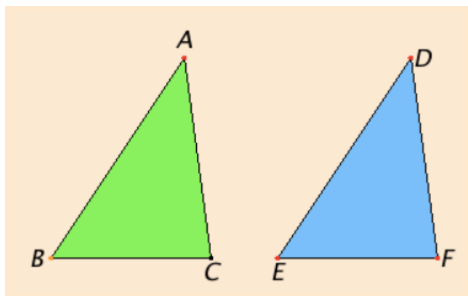
□

At this point, we have a method to mark off a given length from a larger length, even though all we have is a collapsing compass. Therefore, going forward, we can act as if we have a standard compass, that holds its setting after being lifted from the paper.

We also have the means to an important *trichotomy*. Comparing two line segments, one of three things must be true: either the first is smaller than the second, they are equal, or the second is smaller than the first.

Prop. I.4

If two triangles have two sides equal to two sides respectively, and have the angles contained by the equal straight lines equal, then they also have the base equal to the base, the triangle equals the triangle, and the remaining angles equal the remaining angles respectively, namely those opposite the equal sides.



This is not a construction, unlike the previous three propositions. It is a method for proving congruence (equality) of two triangles

$$\triangle ABC \cong \triangle DEF$$

Elsewhere in this book we would call the method SAS or *side angle side*. Given that $AB = DE$ and $AC = DF$ and that the angles between them at the vertices A and D are also equal, the two triangles are congruent: all three angles and all three sides are equal.

This (P I.4) is a proof that SAS is correct.

The proof is by superposition. The facts establish the positions of the points B and C, which determines AB and so the angles at vertices B and C.

Euclid says that if we lift up $\triangle ABC$ and lay it on top of $\triangle DEF$ then B coincides with E and C coincides with F so $BC = EF$.

$\square$

This seems perhaps a little shaky logically, and it's not a method of proof that Euclid uses much.

But one might instead have taken this proposition as a postulate. The source, above, says that David Hilbert claims that under the hypotheses of the proposition it is true that the two base angles are equal, and then proves that the bases are equal.

We have used SAS to prove SSS, that all three sides are equal.

Prop. I.5

In isosceles triangles the angles at the base equal one another, and, if the equal straight lines are produced further, then the angles under the base equal one another.

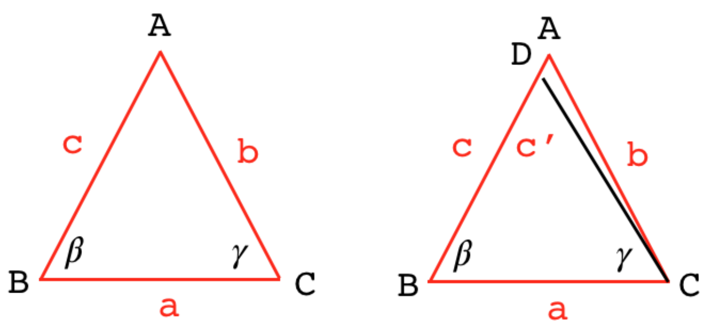
We proved this theorem in the previous chapter on isosceles triangles.

Euclid's proof of the converse is short and introduces the method of contradiction, or *reductio ad absurdum*. That is the next proposition.

Prop. I.6

If in a triangle two angles equal one another, then the sides opposite the equal angles also equal one another.

Suppose we have $\triangle ABC$ with equal angles $\beta = \gamma$ at the base (left panel).



We will assume that the two sides b and c are not equal. Then one of them is greater. Let c be greater, then cut off b from c at point D such that the new length $c' = b$.

The new triangle has sides c' and a, which flank angle β, while for the original we have side b and side a flanking γ. But we constructed $c' = b$, are given that $\beta = \gamma$, and the side a is common.

Therefore the $\triangle DBC \cong \triangle ACB$ by SAS.

But this means that the less equals the greater, which is absurd.

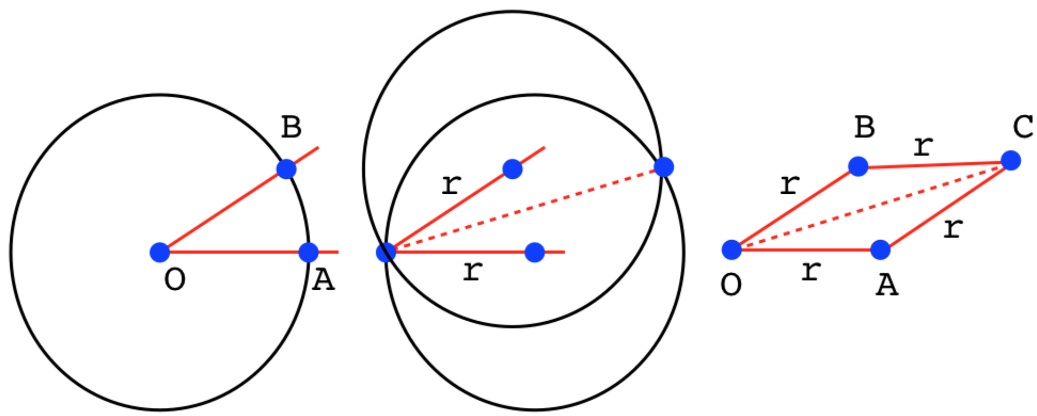
Therefore c cannot be unequal to b. It therefore equals it.

Our original assumption that b does not equal c must be false.

□

Prop. I.9

To bisect a given angle.



As radii of a circle on center O, we first find points A and B equidistant from O (left panel). Let that distance be r.

As radii of circles on the centers A and B that pass through O (so the radius is equal to r), we find C equidistant from A and B (middle panel), with radius also equal to r.

Thus, $OA = OB = AC = BC$ (right panel).

So $\triangle OAC \cong \triangle OBC$.

Therefore $\angle BOC$ is equal to $\angle AOC$ and the given angle is bisected.

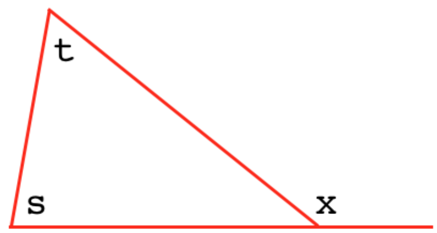
□

We will do three more. They are short, sweet and powerful.

In these examples, we use letters late in the alphabet (s, t, u, v) for angles, while a, b, c are labels for sides.

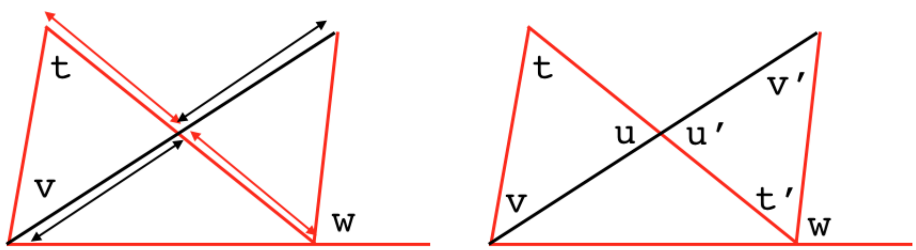
Prop. I.16

In any triangle, if one of the sides is produced (extended), then the exterior angle is greater than either of the interior and opposite angles.



The claim is that the exterior angle x is greater than either of the interior angles: s or t.

Find the midpoint of the side opposite s and draw the indicated line segment (below), so that the two segments marked with black arrows are equal, as well as the segments marked with red arrows.



By SAS and the vertical angle theorem, the two smaller triangles to the left and right are congruent, as indicated in the right panel by the labels on the angles: $t = t', u = u', v = v'$.

The original external angle x is seen to be composed of $t' + w$, that is

$$x = t' + w$$

so clearly (the whole is greater than its parts):

$$x > t'$$

but since $t = t'$:

$$x > t$$

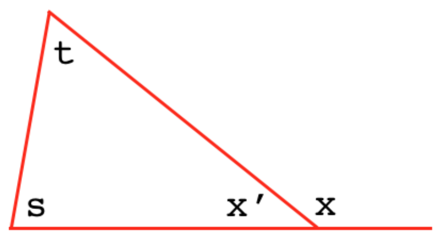
We can make a similar construction and proof for angle s.

The exterior angle is greater than either of the interior and opposite angles.

□

external angle theorem

The external angle theorem is extremely useful, so let's take a break from Euclid and prove it now. The proof is simple for us:



As supplementary angles, $x + x' = 180$ degrees. As the three angles of a triangle, $s + t + x' = 180$ degrees as well. Things equal to the same thing are equal to each other:

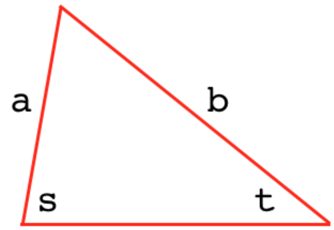
$$x + x' = s + t + x'$$

$$x = s + t$$

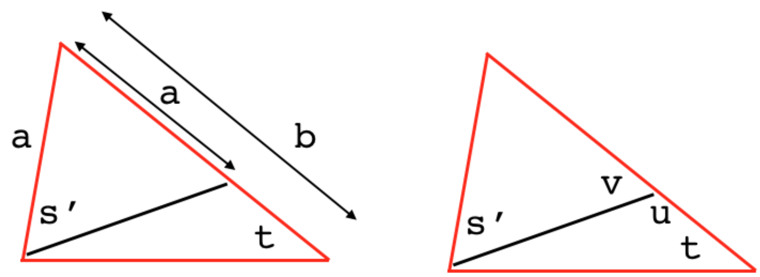
The question of why Euclid doesn't use supplementary angles here is complicated. For now it is enough to say that he just doesn't.

Prop. I.18

In any triangle, a greater side is opposite a greater angle.



Given $b > a$, mark off a on b.



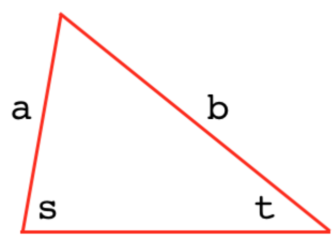
By the external angle theorem (I.16)

$$v > t$$

But $v = s'$ (by isosceles $\triangle$, $I.5$) so

$$s' > t$$

And since $s > s'$



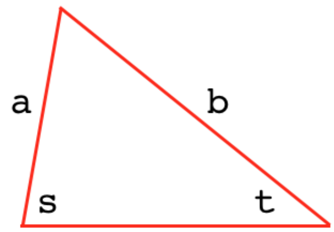
$$s > t$$

$\square$

We get the converse almost for free.

Prop. I.19

In any triangle, a greater angle is opposite a greater side.



We are given $s > t$ and want to prove $a < b$. We proceed by considering the other possibilities.

It cannot be that $a = b$ because then $s = t$ by isosceles $\triangle$ ($I.5$), but we are given $s > t$.

So then suppose $a > b$. By the previous proposition ($I.18$), we would have that $t > s$. But this is again contrary to what we were given. Hence $b > a$.

$\square$

We have made use of the trichotomy from before, that there are only three possibilities:

$$a < b, \qquad a > b, \qquad a = b$$

This applies to line segments and angles as well as many other things.

This is enough of the *Elements* to give us a good taste of the basics of Greek geometry of lines and triangles, and methods of proof. There is more to come: Pythagoras, and circles with their arcs and tangents.

Part IV

Pythagoras

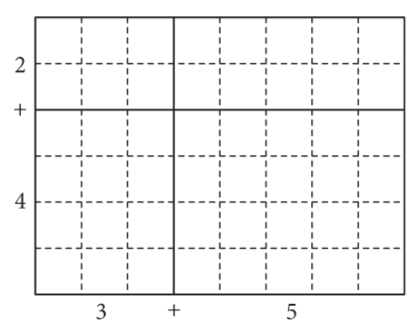
Chapter 16

Area

One aspect of calculus will be to determine the area of figures in the plane, particularly figures bounded by curves, as well as volumes in space. This is the magic of calculus, that we can make curves conform to rectilinear concepts of area and volume.

Since this introductory section is about Euclidean geometry, let's just say a few words about the area of a triangle. But we'll start with the rectangle.

To find the area of a rectangle, we must first fix a unit length. Then multiply the width by the height.



This particular figure (from Lockhart) shows the distributive law in action:

$$(3+5)\cdot(4+2)$$

$$= 3\cdot 4 + 3\cdot 2 + 5\cdot 4 + 5\cdot 2$$

$$= 48$$

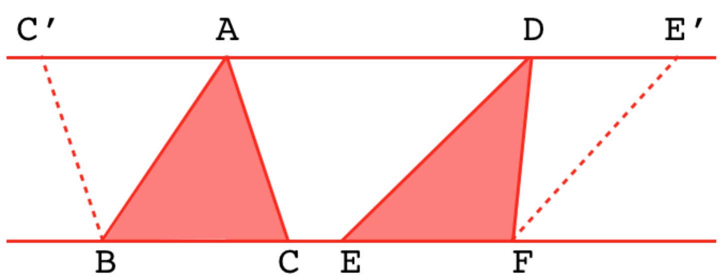
Any combination of numbers that add up to 8, times any combination of numbers that add up to 6, gives the same result.

The next figure is a parallelogram, a four-sided figure whose two pairs opposite sides are parallel (left panel). As a consequence of the theorems we saw previously, the opposing angles are equal, and the adjacent angles add up to 180 degrees.

To find the area, we cut off a right triangle from the left and re-attach it on the right. The angles add up to form a straight line along the base and a right triangle at the upper right. The area is clearly $h \times b$.

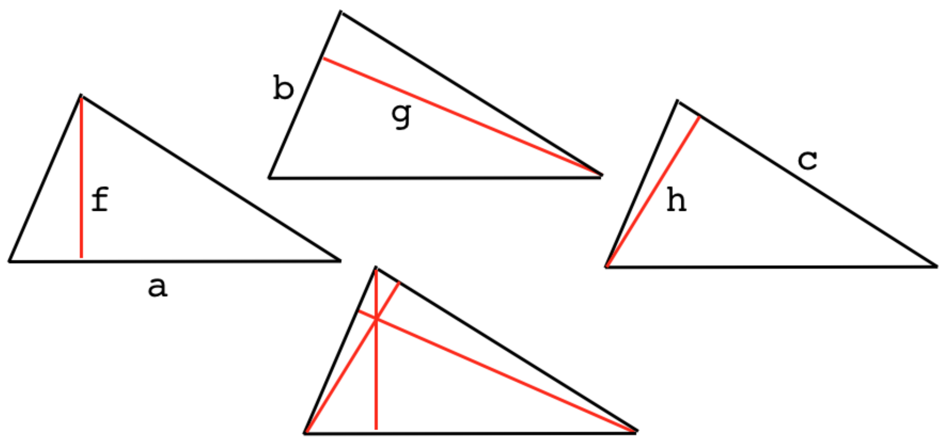
What about triangles? Previously, we talked about an approach where any triangle can be divided into two right triangles.

Alternatively, any triangle can be turned into a parallelogram, by attaching a rotated image of itself, like this:



It is easy to show that $\triangle ABC \cong \triangle ABC'$ and $\triangle DEF \cong \triangle DE'F$. For example, by construction AC is parallel to BC' so $\angle ACB = \angle AC'B$. We'll leave it to you to complete the proof.

An acute triangle is on the left and an obtuse triangle on the right. Since the area of each triangle is one-half that of its corresponding parallelogram (because we added the same area to make the parallelogram), the area of a triangle is one-half the base times the height.



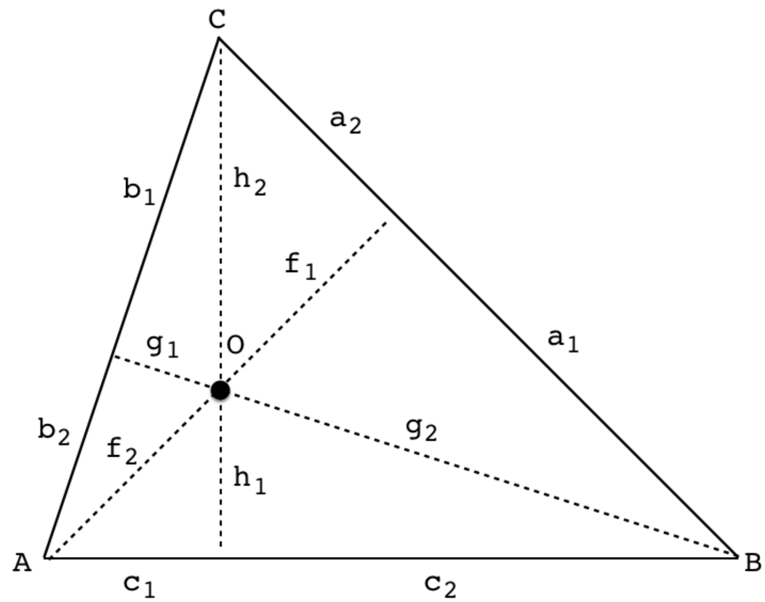
In the figure above, the area is

$$A = \frac{1}{2}\ af = \frac{1}{2}\ bg = \frac{1}{2}\ ch$$

We can choose any side of the triangle to be the base and then multiply $1/2\cdot$ base $\cdot$ height to get the area. We must always get the same answer!

If you accept the argument about the parallelogram above, it must be true, because the area of the triangle has to be the same no matter how you calculate it.

Here's a proof by counting up the area of smaller triangles:



In $\triangle ABC$ with sides a, b, c, drop the three altitudes from each of the three vertices to form right angles on the opposing sides. Ceva's theorem says that these altitudes cross at a single point (we will prove this later). Label the parts of the sides and the altitudes as shown in the diagram.

The area of the whole $\triangle ABC$ is equal to the sum

$$\triangle BOC + \triangle AOC + \triangle AOB$$

Using the rule, *twice* the area is

$$2A = af_1 + bg_1 + ch_1$$

But each of these smaller areas can be computed in different ways. In particular $\triangle BOC$ can be viewed as having base g_2 and height b_1, while $\triangle AOB$ can be viewed as having base b_2 and height g_2, so (twice) the total area is also

$$2A = b_1 g_2 + b_2 g_2 + bg_1$$

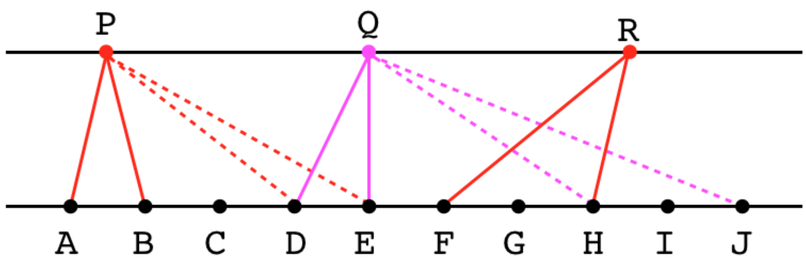
$$= bg_2 + bg_1 = bg$$

Similar calculations can be carried out for the other two sides. Hence the area is the same regardless of which side is chosen as the base.

□

A corollary is that all triangles with the same base and height have the same area.

Draw two parallel lines. Mark off equal distances between adjacent points A through J on the bottom. Now pick any point on the top and draw the triangle with two *equidistant* points on the bottom. Any other triangle drawn with an equal base has the same area.



In this figure the areas of $\triangle PAB$, $\triangle PDE$, and QDE are equal, as are $\triangle QHJ$ and $\triangle RFH$. Further, the latter two have twice the area of any of the first three.

platonic solids

https://en.wikipedia.org/wiki/Platonic_solid

> In three-dimensional space, a Platonic solid is a regular, convex polyhedron. It is constructed by congruent (identical in shape and size) regular (all angles equal and all sides equal) polygonal faces with the same number of faces meeting at each vertex. Five solids meet these criteria.

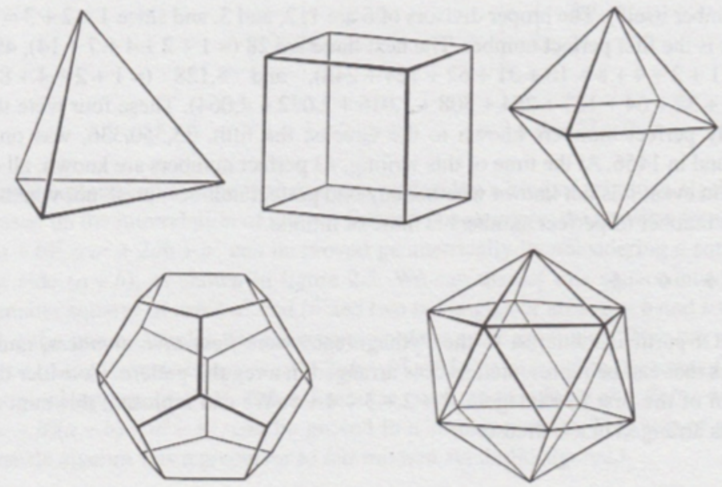


Figure 2.3. The five regular or Platonic solids

These are: (i) tetrahedron, (ii) cube, (iii) octagon, (iv) dodecagon, and (v) icosahedron.

There is a wonderful, simple proof that there are only five of them. Any solid requires at least three sides meeting at each vertex, otherwise the joint between two sides can just flap, like a hinge. Furthermore, the total of all the vertex angles added up must be less than 360 degrees, since otherwise the figure would be planar, not 3-dimensional.

So, three equilateral triangles total $60 \times 3 = 180$, four total $60 \times 4 = 240$ and five total $60 \times 5 = 300$. Six would be a hexagon lying in the plane. Three squares total $90 \times 3 = 270$, while four give a square array in the plane. Finally, three pentagons give $108 \times 3 = 324$. And that's it. Three hexagons would give $120 \times 3 = 360$, which gives an array in the plane.

Proving that all the angles and side lengths come out correctly, so that the possible solids actually can be constructed is another matter, however. Euclid devotes book XIII of *The Elements* to this:

https://mathcs.clarku.edu/~djoyce/elements/bookXIII/bookXIII.html#props

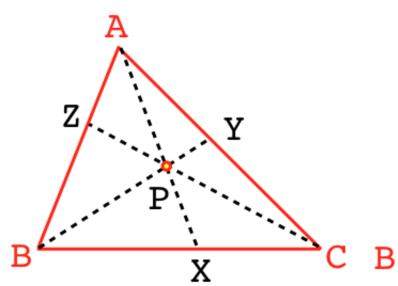
Chapter 17

Ceva's theorem

There are some special points in triangles including orthocenters, which we've already mentioned, but also circumcenters, incenters and centroids.

They have the common feature that three lines are drawn crossing at a single point. We need to establish the conditions under which this assertion is true.

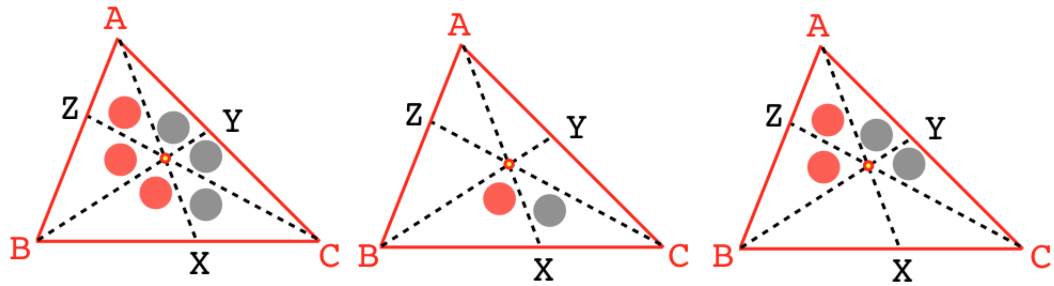
We begin with the triangle shown below, picking a point P to be *any point* inside the triangle. Now draw line segments from each vertex through P and extend them to the opposing side.



Since P can be anywhere, the ratio can be anything. Let's call it x.

$$\frac{BX}{XC} = x$$

Line AX divides the whole triangle into two parts.

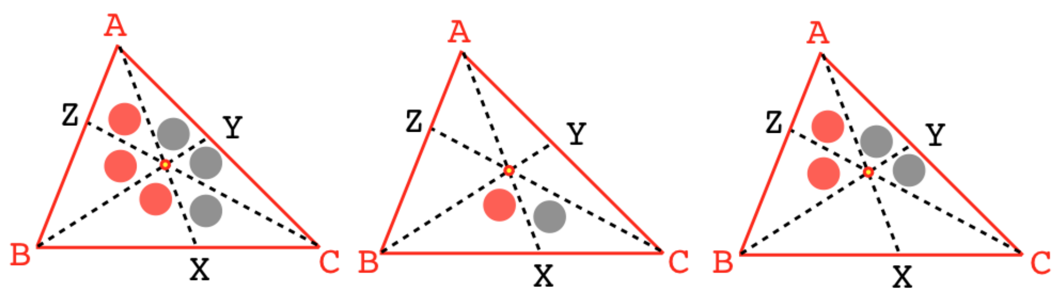


We know that the area of $\triangle ABX$ is in the same proportion to the area of $\triangle AXC$ as x, because they share the same height, while x is the ratio of their bases.

$$BX = x \cdot XC$$
$$A_{ABX} = \frac{1}{2}h \cdot BX = \frac{1}{2}hx \cdot XC = xA_{AXC}$$

Now consider the lower pair of triangles $\triangle BPX$ and $\triangle CPX$

These two also have their areas in the ratio x, for the same reason.



By subtraction, $\triangle ABP$ and $\triangle ACP$ also have ratio x.

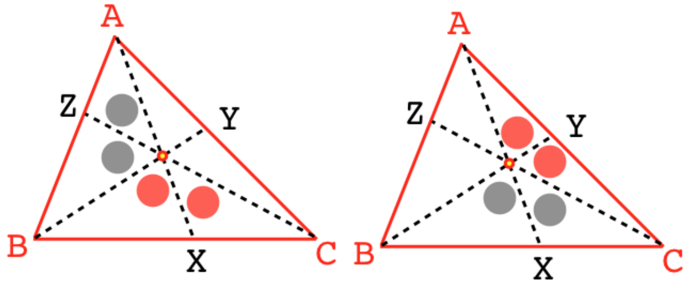
So, altogether, we have that

$$\frac{BX}{XC} = \frac{|ABX|}{|ACX|} = \frac{|BPX|}{|CPX|} = \frac{|ABP|}{|ACP|} = x$$

more sides

By the same reasoning, if $y = CY/YA$

$$\frac{|BCP|}{|ABP|} = y$$



and if $z = AZ/ZB$

$$\frac{|ACP|}{|BCP|} = z$$

Then

$$xyz = \frac{|ABP|}{|ACP|} \frac{|BCP|}{|ABP|} \frac{|ACP|}{|BCP|}$$

But all terms cancel, so

$$xyz = 1$$

And this is of course true not just for the areas but for the original line segments

$$xyz = \frac{BX}{XC} \frac{CY}{YA} \frac{AZ}{ZB} = 1$$

$\square$

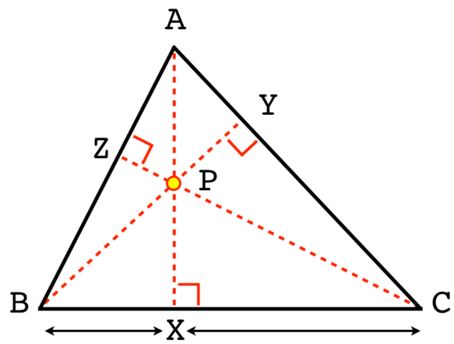
This proof also works in reverse,

$$xyz = 1 \iff \text{3 lines cross at point P}$$

We will just assume that part.

orthocenter

So now, for this triangle



if α is the angle at vertex A and so on, then for example,

$$BX = AB\ cos\beta$$

and

$$\frac{BX}{XC} = \frac{AB\ cos\beta}{AC\ cos\gamma}$$
$$\frac{CY}{YA} = \frac{BC\ cos\gamma}{AB\ cos\alpha}$$
$$\frac{AZ}{ZB} = \frac{AC\ cos\alpha}{BC\ cos\beta}$$

When we construct this ratio, all the terms cancel.

$$\frac{ABcos\beta}{AC\ cos\gamma}\ \frac{BC\ cos\gamma}{AB\ cos\alpha}\ \frac{AC\ cos\alpha}{BC\ cos\beta} = 1$$

which means that

$$\frac{BX}{XC}\ \frac{CY}{YA}\ \frac{AZ}{ZB} = 1$$

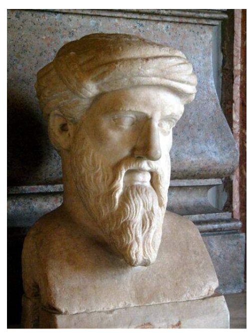
Therefore, the 3 altitudes all cross at a single point. That point is the orthocenter, and this is a proof that it exists.

□

Chapter 18

Pythagoras

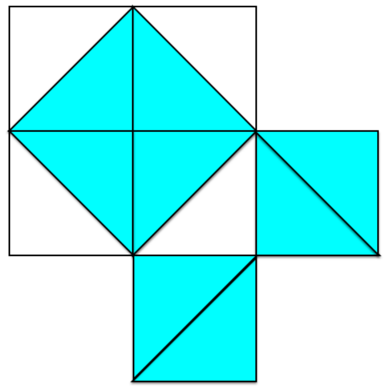
The most famous theorem of Greek geometry is also the most useful in Calculus.



Pythagoras (c.570-c.495 BC) was much younger than Thales but may have encountered him as a young man. Like many other Greek mathematicians, Pythagoras was not from the mainland, but from one of the islands, in his case, Samos, which is not far from Miletus, where Thales lived.

Pythagoras was famous as a philosopher as well as a mathematician. In fact, he founded a famous "school" and it is not sure now which of the theorems developed by this school are due to Pythagoras, and which to his disciples. It is not even clear whether the Pythagorean theorem, as we know it, was known to Pythagoras.

However, it's pretty certain that they knew something. The $3, 4, 5$ right triangle and many other Pythagorean triples (see below) had been known for a thousand years (since 1500 BC). Here is a special case, easily proved, for an isosceles right triangle.

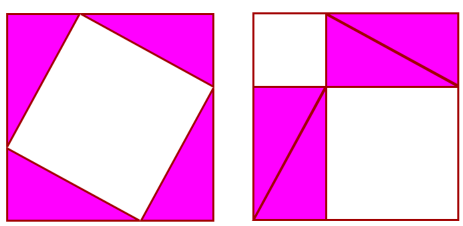


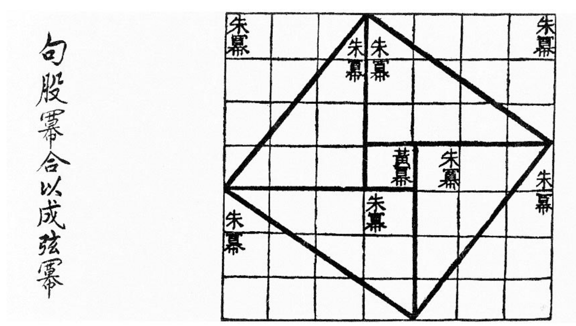
The area of the square on the hypotenuse is equal to twice the area on each side.

There are literally hundreds of proofs of the general theorem, that if a and b are the shorter sides of a right triangle and c is the hypotenuse, then

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

This one is sometimes called the "Chinese proof." I can easily imagine proceeding from the figure above to this one by simply rotating the inner square.





It really needs no explanation, but ..

In the left panel we have a large square box that contains within it a white square, whose side is also the hypotenuse of the four identical right triangles contained inside. Altogether the four triangles plus the white area add up to the total.

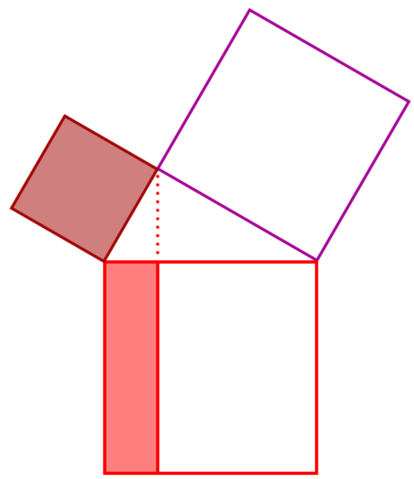
We simply rearrange the triangles. Now we evidently have the same area left over from the four triangles, because they still have the same area and the surrounding box has not changed.

But clearly, now the white area is the sum of the squares on the second and third sides of the triangles. Hence the two white squares on the right are equal in area to the large white square on the left. □

This proof is contained in the Chinese text Zhoubi Suanjing (right panel, above).

https://en.wikipedia.org/wiki/Zhoubi_Suanjing

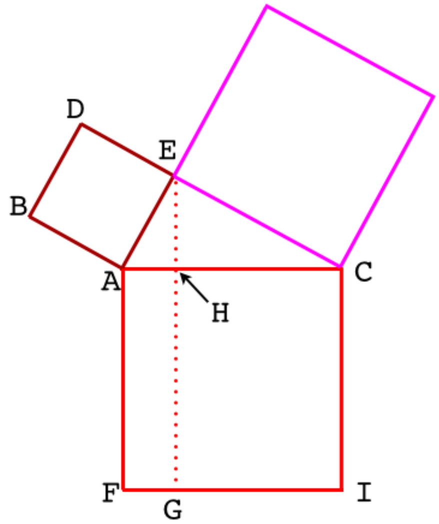
Euclid's proof



My favorite proof relies on the construction above (Euclid $I.47$, sometimes called the "bridal chair" or the "windmill"), where the central triangle is a right triangle, and the other constructions are squares. It is a bit more detailed, but it is a gem of a proof, from Euclid, which is a justification for including it.

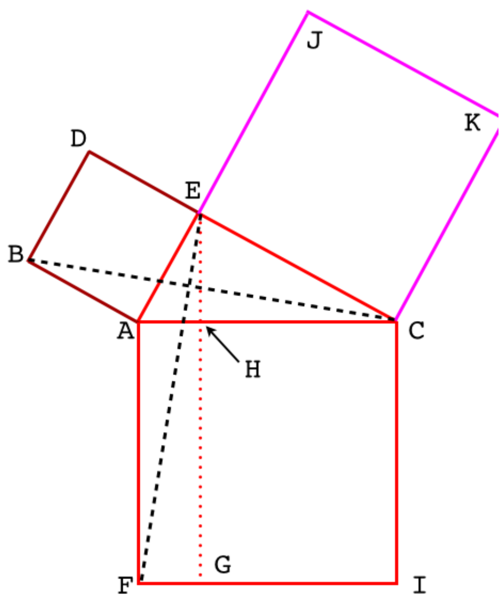
What we will show is that the part of the large square in red is equal in area to the entire small square, in maroon.

We label some points as shown:



First, drop a vertical line EHG, constructing the rectangle $AFGH$.

Finally, sketch dotted lines for the long sides of two triangles:



The crucial point is this: we will show that triangle ΔABC is congruent to triangle ΔAEF.

Use side-angle-side (SAS). The two sets of sides are evidently equal

$$AB = AE, \quad AC = AF$$

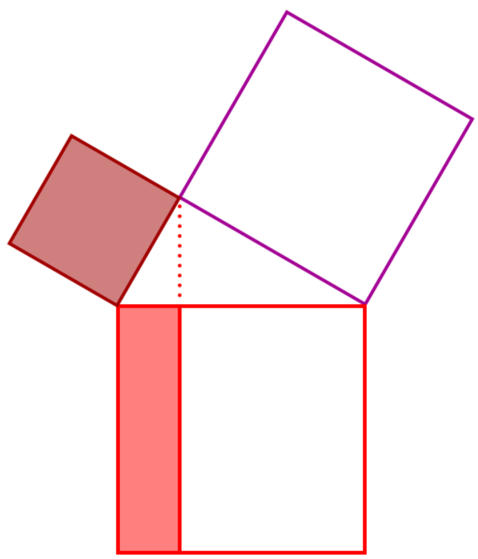
because these are given as sides of two squares.

What about the included angle? The angles $\angle BAC$ and $\angle EAF$ each contain a right angle plus the shared angle $\angle EAC$. So they are themselves equal, and thus we have proved SAS and thus the congruence relationship:

$$\Delta ABC = \Delta AEF$$

The next part of the proof is to tilt triangle ΔABC to the left and see that it has base AB and altitude AE so its area is one-half that of the small square $ABDE$. On the other hand triangle ΔAEF has base AF and altitude AH (as well as FG) so its area is one-half that of the rectangle $AFGH$.

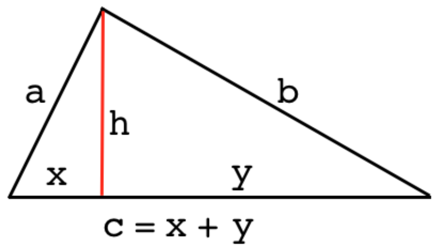
Hence we have proved that the two colored areas in this figure are equal:



Finally, we could proceed to do the same thing on the right side of the figure, but we just appeal to symmetry. All the equivalent relationships will hold.

□

There are several hundred proofs of the Pythagorean theorem. Many of them are algebraic. Here is a classic:



We know that when an altitude is drawn in a right triangle, the two resulting right triangles are similar (use complementary angles if you need to convince yourself again). So we have equal ratios of sides. Here are two sets:

hypotenuse to short side

$$\frac{a}{x} = \frac{b}{h} = \frac{c}{a}$$

hypotenuse to long side

$$\frac{a}{h} = \frac{b}{y} = \frac{c}{b}$$

From the first

$$a^2 = cx$$

From the second

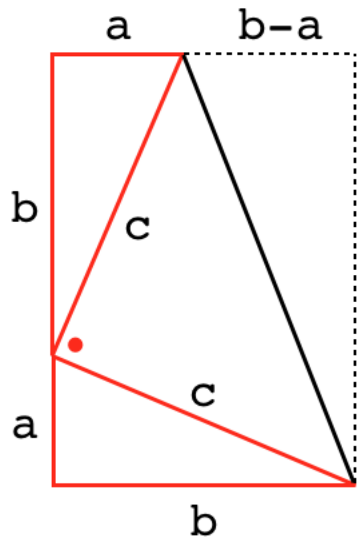
$$b^2 = cy$$

Just add

$$a^2 + b^2 = cx + cy$$
$$= c(x + y) = c^2$$

Garfield

There is one by a future President of the United States, James A. Garfield. (He was a congressman at the time).



Draw a right triangle and a rotated copy as shown. The angles opposite sides a and b are complementary angles. So the angle marked with a dot is a right angle, and the triangle with sides labeled c is a right triangle.

The area of the quadrilateral is the product of the side $(a+b)$ and the *average* of a and b (top and bottom). This can be seen intuitively. The halfway point of the solid red line has horizontal dimension $(a+b)/2$. Hence

$$A = (a+b)\cdot\frac{1}{2}(a+b)$$

If you're worried about that argument, just subtract the area of the triangle with two dotted sides from the quadrilateral that includes it:

$$A = (a+b)b - \frac{(a+b)(b-a)}{2}$$

$$= (a+b)(b - \frac{b}{2} + \frac{a}{2})$$

$$= (a+b)\cdot\frac{1}{2}(a+b)$$

which is just what we said. So now:

$$= \frac{a^2}{2} + ab + \frac{b^2}{2}$$

But we can also calculate the area of the quadrilateral as the sum of the three triangles:

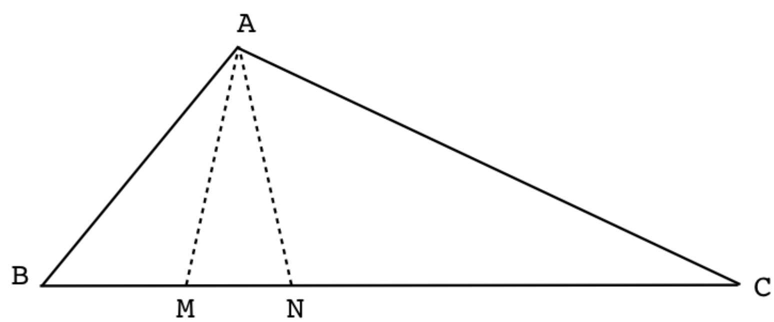
$$A = \frac{ab}{2} + \frac{ab}{2} + \frac{c^2}{2}$$

Equate the two and the result follows almost immediately.

□

Corollary

There are several important corollaries of the Pythagorean theorem. We'll derive one later called the law of cosines. Here is another from the Islamic geometer Ibn Quorra, who brought algebraic techniques, shunned by the Greeks, to geometry.



Let $\triangle ABC$ be *any* triangle (here it is obtuse). Draw AM and AN so that the new angles $\angle AMB$ and $\angle ANC$ are equal to $\angle A$. The corresponding triangles are similar to the original, because they share the angle of measure A plus one other from the original triangle.

Then

$$BM : AB = AB : BC$$

Thus, $AB^2 = BM \times BC$. Similarly

$$NC : AC = AC : BC$$

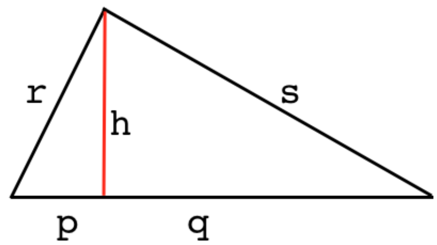
So $AC^2 = NC \times BC$ Therefore

$$\begin{aligned} AB^2 + AC^2 &= BM \times BC + NC \times BC \\ &= (BM + NC) \times BC \end{aligned}$$

In the case where the angle at vertex A is a right angle, then M coincides with N, and $BM + NC = AC$, and this reduces to the Pythagorean theorem.

geometric mean

As a slight detour from calculus, but on the topic of this chapter



We will show that

$$h^2 = pq$$

$$h = \sqrt{pq}$$

That is, h is the geometric mean of these two values p and q.

Proof.

Using the Pythagorean theorem with the two small triangles (also right triangles), we obtain:

$$h^2 + p^2 = r^2$$

$$h^2 + q^2 = s^2$$

Summing

$$2h^2 + p^2 + q^2 = r^2 + s^2$$

Using the theorem with the big triangle:

$$r^2 + s^2 = (p+q)^2$$

$$= p^2 + 2pq + q^2$$

Equating the two expressions for $r^2 + s^2$ we get:

$$2h^2 + p^2 + q^2 = p^2 + 2pq + q^2$$

$$h^2 = pq$$

$$h = \sqrt{pq}$$

Part V

Circles

Chapter 19

Circle and triangle

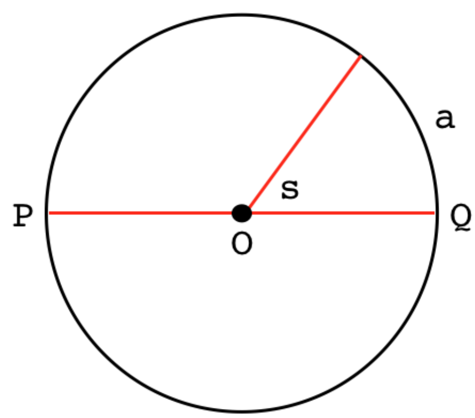
From a previous chapter, Euclid's third postulate was:

• Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center. The tool to do this is a compass.

If the radius is extended so that it cuts the circle at two points, it is called a diameter.

We saw previously that one can construct a line perpendicular to any given line. If that line is constructed perpendicular to the diameter at the point where it meets the circle, the new line is called a tangent line. By definition, the tangent line touches the circle at a single point.

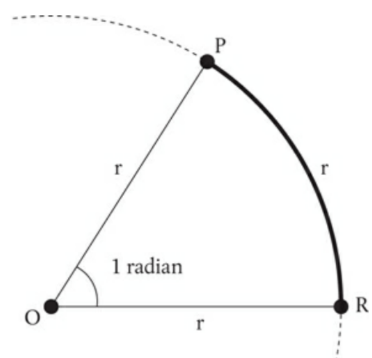
arcs of a circle



In calculus and analytical geometry angles are defined in terms of radians of arc. In the figure above, the angle s is *defined to be equal to* the length of arc a that it sweeps out, or subtends in a unit circle.

For a unit circle with radius = 1, the total circumference is 2π, so the arc swept out by the angle θ, measured in radians, is in the same ratio to 2π as the ratio of the angle's measure in degrees to 360°.

It seems natural then to adopt the arc length as a measure of the angle, where 360° is equal to 2π *radians*, and an angle of 90°, for example, a right angle, is equal to $\pi/2$ radians.



72. Definition of a radian.

One can easily divide 360 by 2π to find that one radian is approximately 57°.

To convert some more measures of angles in degrees to radians:

$$180^\circ = \pi,\ 90^\circ = \frac{\pi}{2}$$

$$60^\circ = \frac{\pi}{3},\ 45^\circ = \frac{\pi}{4},\ 30^\circ = \frac{\pi}{6}$$

Central angle and subtended arc are numerically equal, but remember that they are dimensionally different. Arc is a length, angle is just an angle.

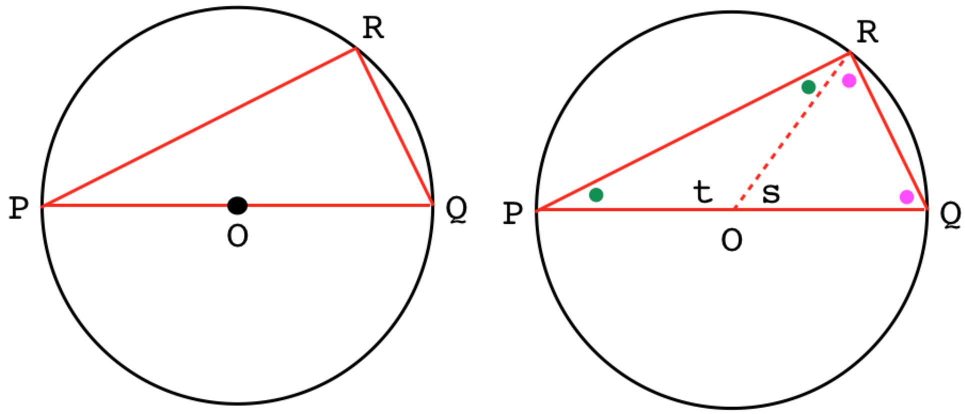
Thales' theorem

In this chapter, we introduce a few more theorems concerning circles, starting with the last of Thales' theorems:

- Any angle inscribed in a semicircle is a right angle.

Think of three points on the circumference of the circle as forming a triangle. If two points are on a diameter of the circle, the angle formed at any arbitrary but distinct third point is always a right angle.

To prove: $\angle PRQ$ is a right angle.

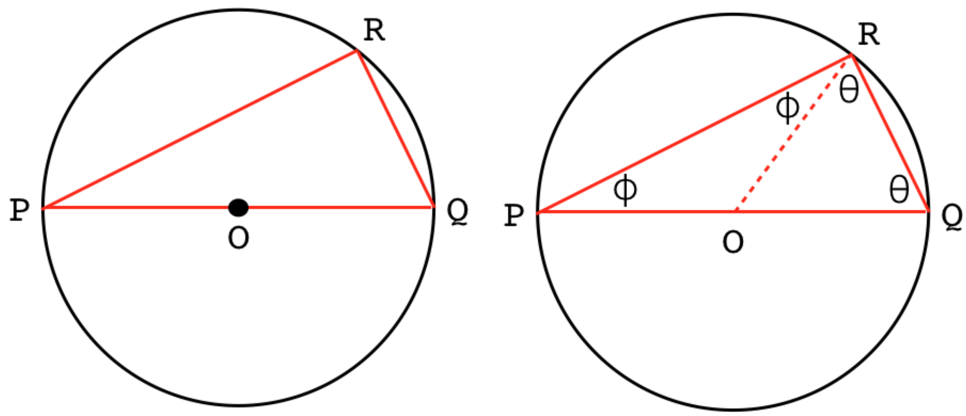


Solution:

Draw the radius OR. Notice that the two smaller triangles produced ($\triangle OPR$ and $\triangle OQR$) are both isosceles, since two of their sides are radii of the circle.

Therefore, in each triangle the two angles marked with dots of the same color are equal (by P $I.5$).

Since $\angle PRQ$ contains one angle of each type, it is equal to one-half the angle sum for the triangle, i.e., $\angle PRQ$ is a right angle.



To restate this: in the figure above, $\angle PRQ = \phi + \theta$. Since the full measure of the triangle is $180° = \pi$ radians, and

$$\phi + \phi + \theta + \theta = \pi$$

it follows that

$$\phi + \theta = \pi/2$$

$\square$

angles on the perimeter

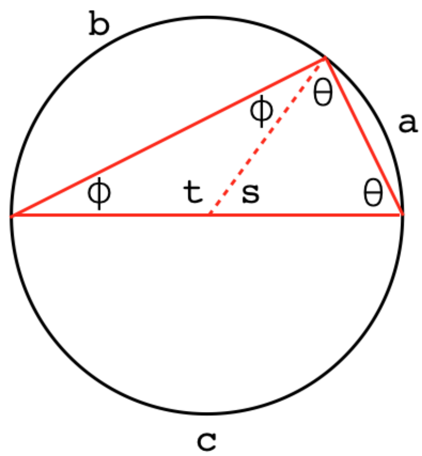
The arc swept out by the right angle $\angle PRQ$ is clearly equal to π, because that arc is one-half of a circle.

But the angles on the perimeter of the circle subtending the very same arc add up to

$$\angle PRQ = \phi + \theta = \pi/2$$

What's going on?

To clarify, let us label the arcs on the circle.



a is the arc swept out by angle s, and a and s have the same measure by definition, although one is a length and the other an angle.

$$s = a$$

By the external angle theorem, we know that

$$s = 2\phi$$

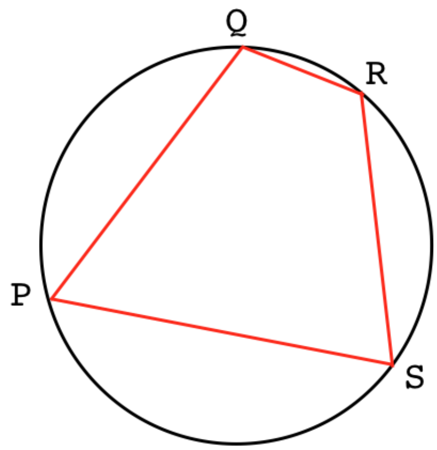
and so conclude that

$$\phi = \frac{s}{2}$$

The angle ϕ lying on the perimeter sweeps out the same arc as s, even though ϕ is one-half of s.

ϕ is farther away from the arc, so we get a bigger arc for the same angular measure. The arc length swept out by an angle on the perimeter is twice the angle's measure in radians.

This leads to a wonderful simple theorem about quadrilaterals.

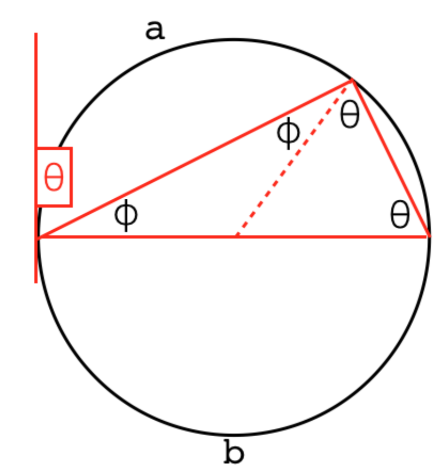


For *any* quadrilateral whose four vertices lie on a circle, the opposing angles are supplementary (they sum to 180°).

Proof: together, opposing angles exactly subtend the whole arc of the circle.

tangent

Consider the chord PR and draw the tangent at P.



The angle between the tangent and the chord equals θ because $\theta + \phi$ is a right angle

Take a chord of the circle, draw the diameter and the tangent. The same rule applies to both angles: one between the chord and the diameter, and the second between the chord and the tangent. The arc is twice the measure of the angle.

geometric mean

We showed in the chapter on the Pythagorean theorem that the altitude of a right triangle is the geometric mean of the two components of the base.

$$h^2 = pq$$
$$h = \sqrt{pq}$$

According to wikipedia:

https://en.wikipedia.org/wiki/Geometric_mean

The fundamental property of the geometric mean is that (letting m be the *geometric mean* here):

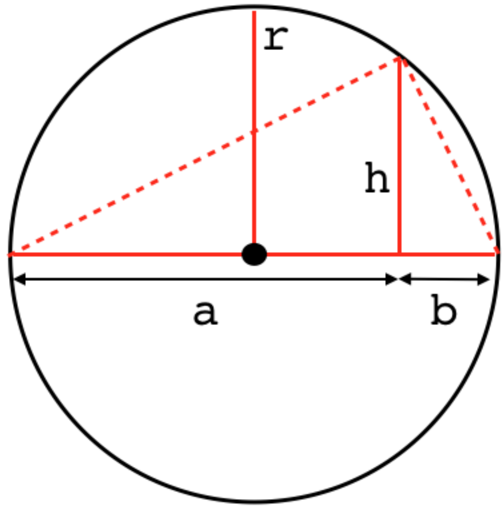
$$m \left[\frac{x_i}{y_i} \right] = \frac{m(x_i)}{m(y_i)}$$

and one consequence is that

> This makes the geometric mean the only correct mean when averaging normalized results; that is, results that are presented as ratios to reference values.

A number of examples are given in the article.

This section is here because originally, there was a proof-without-words that the geometric mean is always less than or equal to the arithmetic mean.



I decided to add some words. A right triangle is inscribed in a semicircle. As just mentioned, the altitude h squared is equal to the product of chord segments (we will prove this geometrically in the next chapter as well).

$$h^2 = ab$$

$$h = \sqrt{ab}$$

But we also have that $a + b = 2r$ and hence

$$r = \frac{a + b}{2}$$

Do you recognize these? The second expression is the arithmetic mean of a and b, while the first is the geometric mean.

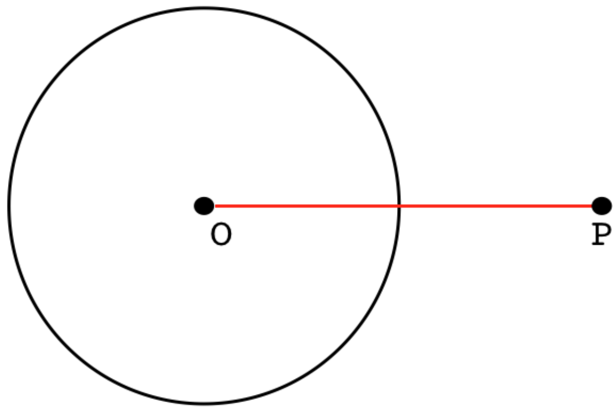
The geometry shows that $h \leq r$ so:

$$\sqrt{ab} \leq \frac{a + b}{2}$$

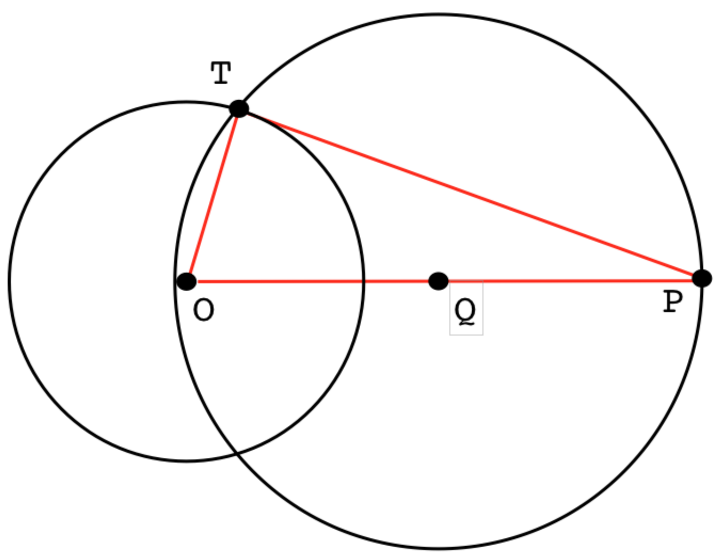
The geometric mean is always less than the arithmetic mean, except when $a = b$, when they are equal (or all of n values are equal).

tangents

Thales theorem provides a way to construct the tangent to a circle passing through any exterior point P.

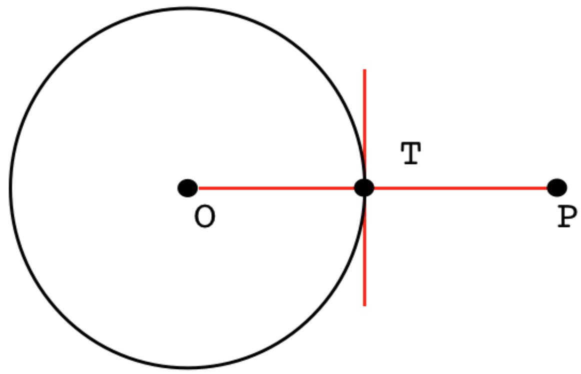


Use OP as the diameter of a circle. Draw the line segment OP and divide it in half by erecting the perpendicular bisector at Q. Use that Q as the center of a new circle. The point T is the intersection of the two circles.



By Thales theorem, $\angle OTP$ is a right angle, and since OT is a radius of the original circle, TP is the tangent at the point T.

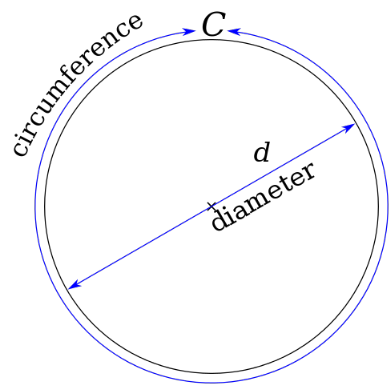
To construct a tangent on a circle at a given point T



Extend OT to P. Construct the perpendicular bisector at T. That is the tangent of the circle.

Chapter 20

Pi is a constant



We began the book with a bold claim: the ratio of the circumference of a circle to its diameter is a constant, independent of the length of the diameter:

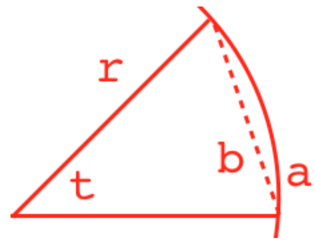
$$\pi = \frac{C}{d} = \frac{C}{2r}$$

We did not prove this theorem at the time but will do so now.

We need the idea of limits, which was introduced previously, and a property of similar triangles. The theorem is: if two triangles are similar, then their sides are proportional to each other.

Consider an arc a of a circle and two radii.

The triangle corresponding to that arc has base b. We can restate Archimedes argument about inscribed polygons by saying that, in the limit, as the inscribed polygon gets very close to being the same as the circle, $b \to a$.

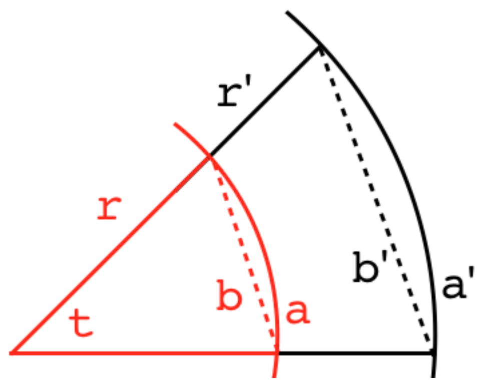


So if there are n pieces ($t = 2\pi/n$), the ratio of the circumference to the arc is just n and we have

$$n = \frac{na}{a} = \frac{C}{a} = \frac{C}{b}$$

The last step is ”in the limit.”

Now draw a larger arc. In the same way $b' \to a'$.



and

$$n = \frac{C'}{b'}$$

since t and n haven't changed:

$$n = \frac{C'}{b'} = \frac{C}{b}$$

But by similar triangles the ratios are equal:

$$\frac{r}{b} = \frac{r'}{b'}$$

so

$$\frac{C'}{C} = \frac{b'}{b} = \frac{r'}{r}$$
$$\frac{C'}{r'} = \frac{C}{r}$$

Suppose for a moment that $C = 2\pi r$ and $C' = 2\pi' r'$ and we don't know how π compares to π':

$$\frac{2\pi' r'}{r'} = \frac{2\pi r}{r}$$

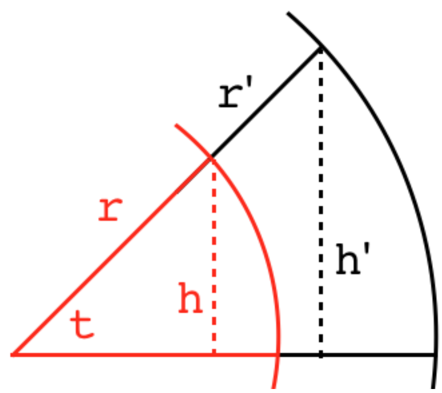
We have that $\pi = \pi'$.

$\square$

Second proof:

Here is a simple variant which assumes something we will prove in the section on sine and cosine. If this is confusing, it can easily be skipped.

Drop the altitude h in each of the two similar triangles. The ratio h/r is equal to $\sin t$, but the arc length is equal to t, measured in radians.



In the limit that $n \to \infty$, the ratio between t and $\sin t = h/r$ is equal to our "special limit":

$$\lim_{n\to\infty} \frac{t}{\sin t} = 1$$

If the ratio to the sine is equal to 1, so is the ratio to its inverse and thus the ratio s/r is constant, which is what we wanted to prove.

$\square$

Chapter 21

Arcs of a circle

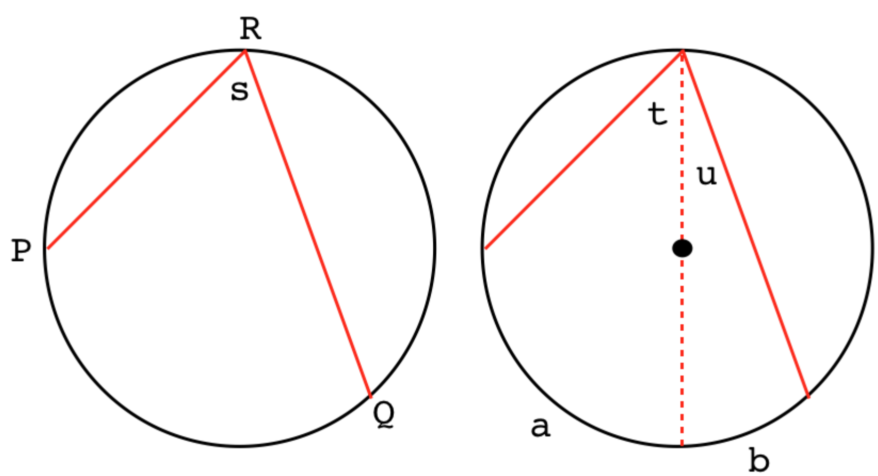
We have previously established some central facts about circles including Thales' theorem about the angle intercepting a half-arc of the circle being a right angle. We will need some of these results later.

Now, we do more. We can generalize the results for all arcs.

arcs encompassing diameter

The examples so far directly contain the diameter in some way.

Consider the arc swept out by the angle s in this figure (left panel).



We can prove that the measure of the angle s is equal to one-half the arc swept out between P and Q.

Draw the diameter (right panel): By our previous work:

$$2t = a, \qquad 2u = b$$

and

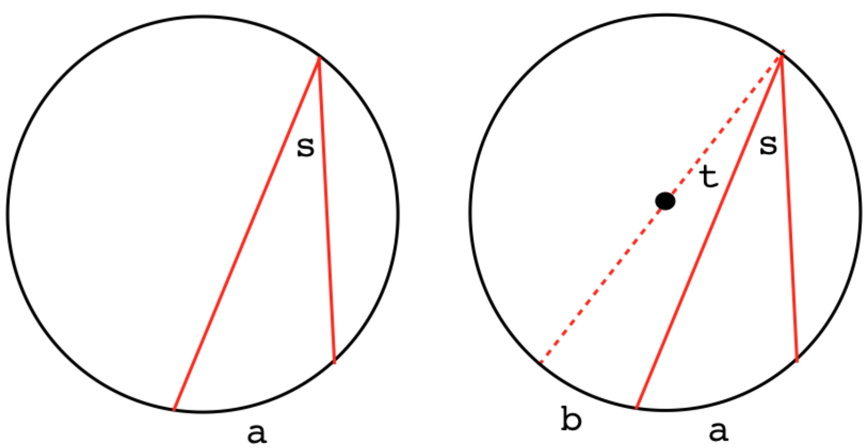
$$s = t + u$$
$$2s = 2t + 2u = a + b$$
$$s = \frac{1}{2}\,(a + b)$$

Thus, we have proved the theorem for

- the special case where the arc is a diameter of the circle
- the general case where the diameter is one chord flanking the arc
- another general case where the arc includes the diameter.

arc without diameter

However, the theorem is true even if the angle does not include the diameter. We use subtraction.



On the right, draw the diameter. There are two arcs which include the diameter: one with angle t and one with angle $s + t$. We obtain the generalized arc with angle s by subtracting the result for t from that for $s + t$.

$$2t = b$$

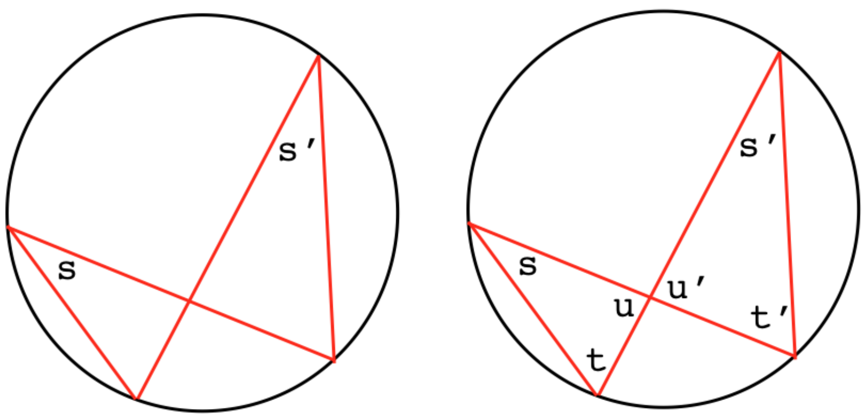
$$2(s + t) = a + b$$

Subtract:

$$2s = a$$

$$s = \frac{1}{2} a$$

As a corollary, any two angles with vertexes (vertices) on the circle that cut off the same arc are equal. In the figure below, $s = s'$ (left panel)



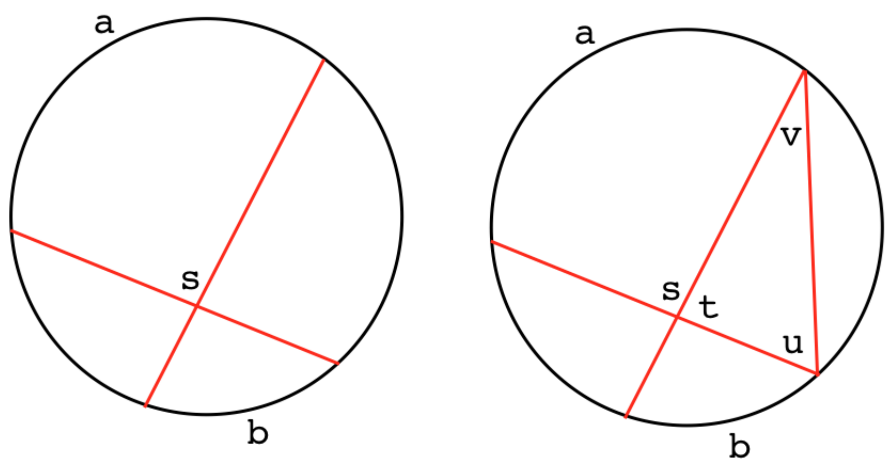
and $t = t'$ (right panel). So $u = u'$ by both the vertical angle theorem and by of the triangle sum theorem. Therefore, the two triangles are similar. We will come back to this.

Intersecting chords

Given two chords, to prove:

$$s = \frac{1}{2}(a + b)$$

s is the average of the two arc lengths.



Solution: Draw a triangle (right panel, above).

$$2v = b$$

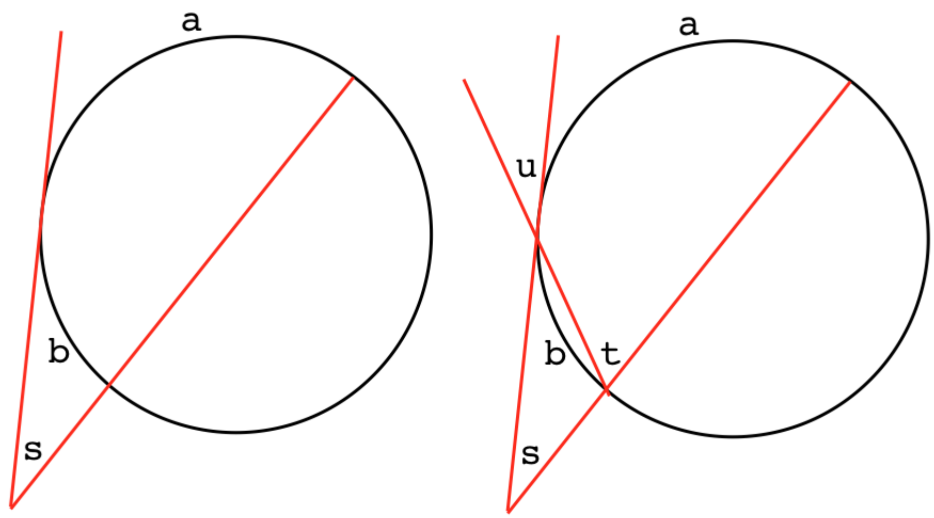
$$2u = a$$

The external angle is the sum of the two opposing interior angles.

$$s = u + v = \frac{1}{2}\ (a + b)$$

Tangent and secant

Rather than having all three points on the circle, one point is now outside. We have the same arc swept out by the endpoints (a), but the included angle is smaller, and there is a new small piece of arc length b.



To prove:

$$s = \frac{1}{2}(a - b)$$

Solution: Draw the triangle. By our previous work:

$$2t = a$$

By the vertical angle theorem the unlabeled angle inside the triangle is equal to u and it subtends arc b so

$$2u = b$$

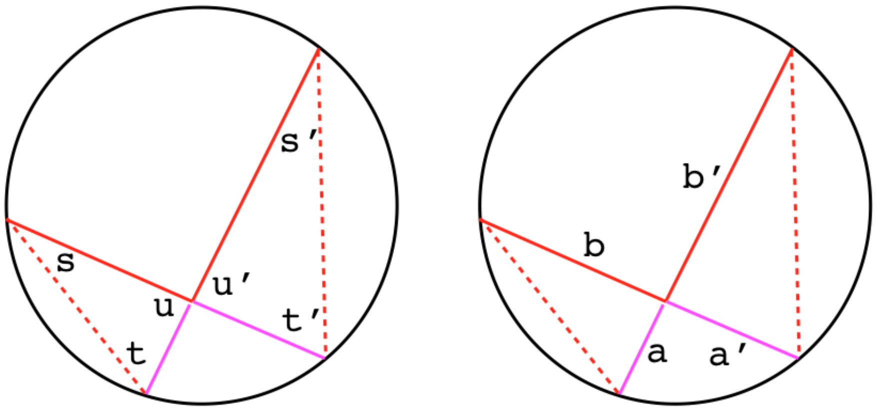
By the exterior angle theorem

$$t = s + u$$

$$s = t - u = \frac{1}{2}\ (a - b)$$

Chord segments

Finally, there is a simple algebraic relationship between chord segments. Draw two chords of the circle.



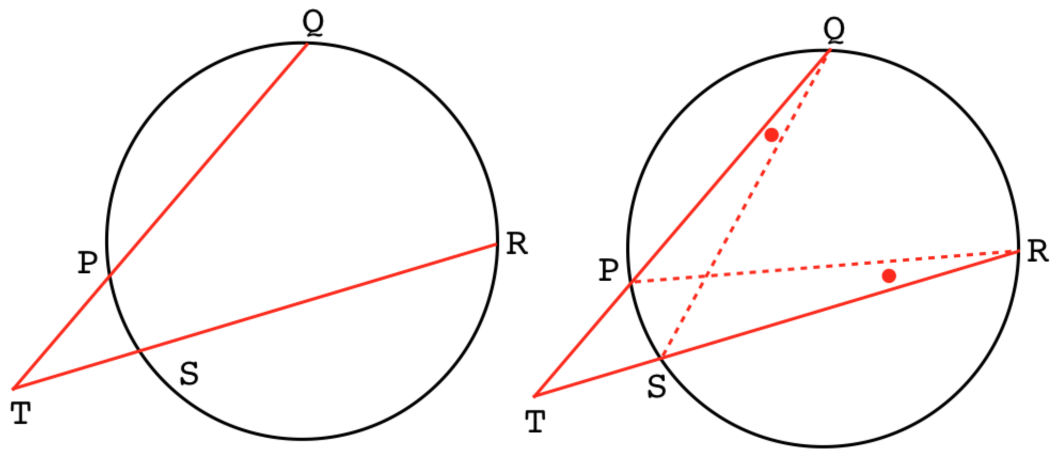
Draw the two triangles. The angles are equal to their primed counterparts ($s = s'$ and $t = t'$ because they subtend equal arcs, while $u = u'$ by the vertical angle theorem).

Therefore, the triangles are similar. Similar sides are indicated by color or style, and labeled with primes in the right panel, where the letters refer to the sides.

$$a/b = a'/b'$$
$$ab' = a'b$$

For any two chords that cross in a circle, the products of the chord segments are equal. We use this later in looking at the spherical cap.

We can prove a similar theorem about chords extended from a circle.



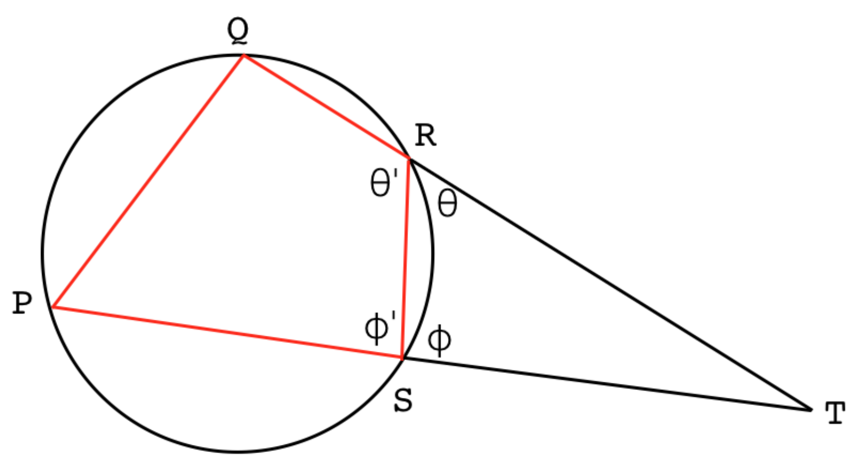
The two angles marked with red dots subtend the same arc, so they are equal. The angle at vertex T is shared, therefore $\triangle QST \cong \triangle PRT$.

By similar triangles we have that

$$\frac{TP}{TS} = \frac{TQ}{TR}$$

$$TP \cdot TR = TS \cdot TQ$$

cyclic quadrilateral



A cyclic quadrilateral is a four-sided figure where all four vertices lie in a circle. Surprisingly, the two triangles in the figure are similar.

To prove: $\triangle PQT \sim \triangle RST$.

The supplementary angle ϕ' subtends arc PQR, therefore ϕ is equal to the angle at vertex Q, which subtends arc PSR.

Similarly, θ' subtends arc QPS, therefore θ is equal to the angle at vertex P.

$\square$

Chapter 22

Eratosthenes

This part of the book is focused on geometry, and we take a look at Eratosthenes in this chapter as an important Greek scholar.

The widely held theory, that the ancient world believed the earth to be flat, is just wrong. People with any level of sophistication not only knew the earth is roughly spherical but also knew its size within a few percent of the true value.

One likely basis is the false story that Columbus had trouble getting financing for his proposed trip to China because everyone thought he would fall off the edge of the earth. This was a tall tale invented by Washington Irving, who also made up several remarkable fables about George Washington.

The real reason the Italians and the Portuguese thought Columbus would fail is that they had a pretty good idea of the size of the spherical earth and thus of the distance to China, while the over-optimistic Columbus believed it was about half the true value. The prospective financiers knew that he was not able to carry the supplies necessary for a trip of this length.

Morris Kline (*Mathematics and the Physical World*) says that the error is due to geographers after Eratosthenes, who reduced the estimated circumference from 24,000 to 17,000 miles.

Eratosthenes

Views of the Greek philosophers on the earth and its sphericity are detailed here https://www.iep.utm.edu/thales/#SH8d

Here is a partial quotation:

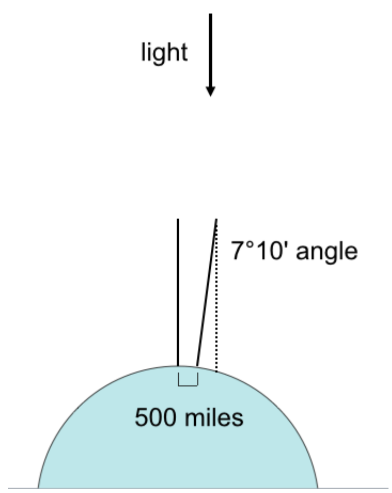
> There are several good reasons to accept that Thales envisaged the earth as spherical. Aristotle used these arguments to support his own view [...] . First is the fact that during a solar eclipse, the shadow caused by the interposition of the earth between the sun and the moon is always convex; therefore the earth must be spherical. In other words, if the earth were a flat disk, the shadow cast during an eclipse would be elliptical. Second, Thales, who is acknowledged as an observer of the heavens, would have observed that stars which are visible in a certain locality may not be visible further to the north or south, a phenomen[on] which could be explained within the understanding of a spherical earth.

https://en.wikipedia.org/wiki/Eratosthenes

Eratosthenes (ca. 276 - 195 BCE) measured the circumference of the earth from this observation: at high noon on June 21st there was no shadow seen at Syene, e.g., allegedly from a stick in the ground. Some people say it was a deep well, where the bottom was illuminated at midday.

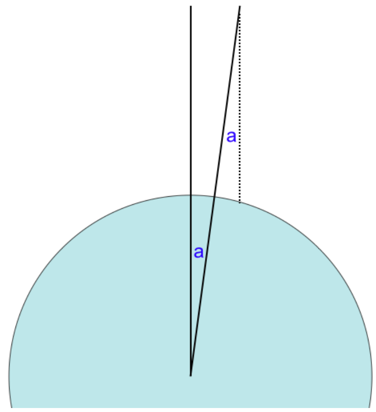
Syene is presently known as Aswan. It is on the Nile about 150 miles upstream of Luxor, which includes the famous site called the Valley of the Kings. At 24.1 degrees north latitude, Aswan or Syene is close enough to having the sun directly overhead on June 21. (The "Tropic of Cancer" is at 23 degrees, 26 minutes north).

This news about the lack of a shadow at Syene reached Alexandria, a famous center of learning of the ancient world. Alexandria lies on the Mediterranean some 500 miles north of Syene, and anyone there who was looking could observe that at high noon on June 21st there *was a shadow*. This shadow Eratosthenes measured to be some 7 degrees and a bit (7 degrees and 10 minutes).



A full 360 degrees divided by 7 degrees and a bit is approximately 50. So we can calculate on this basis that the circumference of the earth is about $50 \times 500 = 25000$ miles. That's pretty close to the correct value.

For this calculation, we assume that the sun's rays are effectively parallel (not a bad assumption given a distance of 93 million miles). Then we just use this:



an application of the alternate-interior-angles theorem.

It is curious how the distance from Alexandria to Syene was calculated.

Kline:

> Camel trains, which usually traveled 100 stadia a day, took 50 days to reach Syene. Hence the distance was 5000 stadia...It is believed that a stadium was 157 meters.

We obtain

$$157 \times 5000 \times 50 = 39,250 \text{ km}$$

That's a much better estimate than a method that relies on camels really deserves.

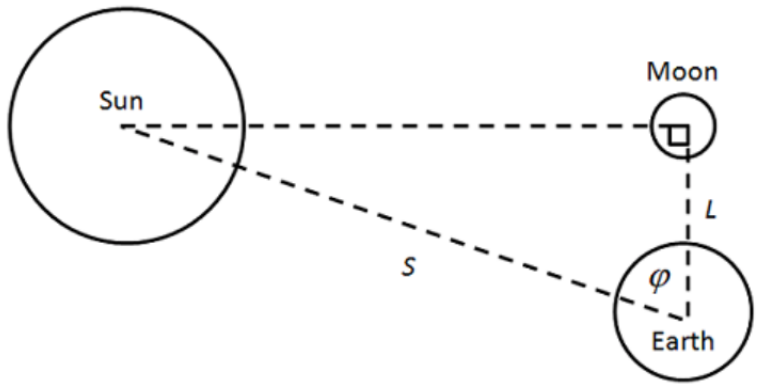
Aristarchus

Aristarchus of Samos (310-230 BCE) wrote a famous book in which he calculated the relative sizes of the sun and the moon and their distances from earth.

One straightforward observation is that the apparent size of the sun and moon in the sky is about the same. This can be seen during a solar eclipse, or observed at any other time by holding a disk up at a fixed distance from the eye, (while taking care to block most of the sun's rays). The value is approximately one-half degree.

Since the distance to the sun is much greater than that to the moon (see below), we can infer that the sun is much larger than the moon.

The central idea of Aristarchus is that, at half moon, the geometry of the three orbs is like this:



In other words, when the phase is half moon and that moon is exactly overhead, the sun has not yet set, but is a bit above the horizon.

If S is the distance to the sun and L is that to the moon, he estimated that

$$18 < \frac{S}{L} < 20$$

with the same ratio for their sizes. Unfortunately, this is not a particularly good estimate. The true value is about 390. Aristarchus obtained a value of 20 for the Earth-Moon distance in Earth radii. The correct value is about 60. Much better estimates were obtained later, by Hipparchus and Ptolemy.

However, Aristarchus made up for this by being the first person to propose a heliocentric theory of the solar system: that the earth and planets rotate around the sun.

https://en.wikipedia.org/wiki/On_the_Sizes_and_Distances_(Aristarchus)

quick estimate

Here is an estimate for the earth-moon distance based on a lunar eclipse.

One measures the time it takes for a complete, total eclipse. From the first shadow of the earth on the moon to the last, that time is about 3 hr. The moon has moved approximately 1 earth diameter in its orbit in that time.

However, we must correct for the fact that the first and last shadows occur on opposite edges of the moon. It was noted that the shape of the eclipse suggests the earth's diameter (at that distance) is about 2.5 moon diameters. So the moon has actually moved $(2.5 + 1.0)/2.5 = 1.4$ earth diameters in the given time. The relevant time becomes 2.14 hr.

Any correction for the true size of the earth's diameter is minimal because the earth-moon system is so far from the source of illumination.

The other piece of information we need is the time for a full revolution, one lunar cycle. This part is tricky. Naively, you'd look for the moon to be in the same place against the fixed stars (27 days, c. 8 hr). This is off because the earth has moved in the meantime — there is a parallax error. As a rough correction, mutliply by 360/330 degrees. The result in hours is 715.

The circumference of the orbit is then

$$715/2.143 = 333$$

earth diameters.

This gives a radius of 53 earth diameters, which is not too far from 60.

Chapter 23

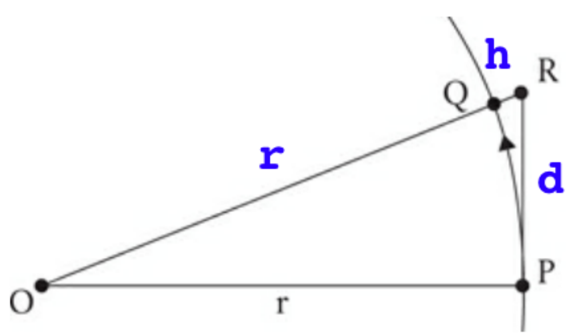
Circular orbits

Pythagoras and Newton

A previous chapter looked in detail at Pythagoras' Theorem, which is used incessantly from here on out. Here, we explore one use of the Pythagorean theorem and provide a taste of orbital mechanics, which is a particular focus of calculus. Newton made early calculations similar to these, which increased confidence about his famous inverse-square law and inspired the mathematics that led to the explanation of elliptical orbits.

Although the orbits of the planets around the sun are ellipses, they are very nearly circular and we will make that approximation for what follows here.

We use the Pythagorean Theorem to make another approximation. Using r for the (fixed) radius of the orbit for the moment, because the construction has capital letters for the points, including the symbol R:



$$r^2 + d^2 = (r+h)^2 = r^2 + 2rh + h^2$$
$$d^2 = 2rh + h^2$$

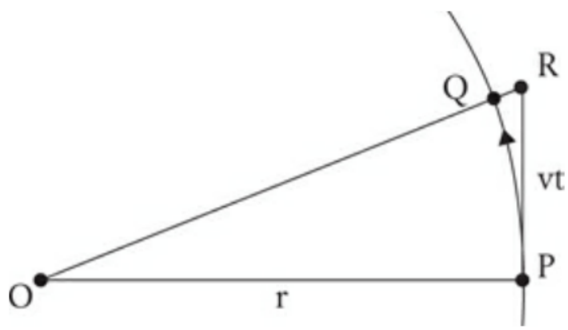
If $h << r$ then we can ignore the very small quantity h^2 and obtain

$$d^2 = 2rh$$
$$r = \frac{d^2}{2h}, \quad h = \frac{d^2}{2r}$$

If the planet were not accelerated, then it would move from P to R, a distance d, and this is equal to the velocity $\times$ time:

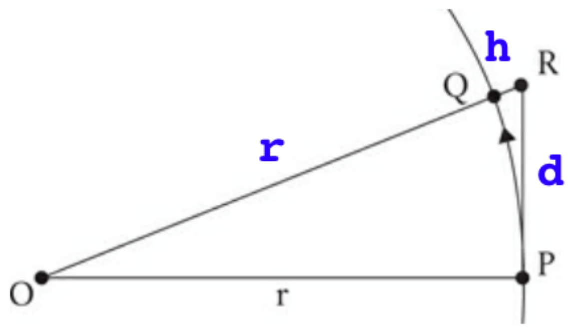
$$d = vt$$

At this point, we use an idea from calculus. *For a small enough segment of the orbit*, this distance PR is the same as the arc length PQ.



So we substitute for $d^2 = (vt)^2$ into the equation from above

$$h = \frac{d^2}{2r} \approx \frac{(vt)^2}{2r}$$



Also, for a small enough part of the orbit (again), h and d are perpendicular to each other as well.

At this point we use the additional assumption that the force is directed toward the sun. We might say that the distance *fallen* by the planet in this short time is h.

By the standard equation of motion, under gravitational acceleration g is related to h and the time t by this equation:

$$h = \frac{1}{2}gt^2$$

We combine the two different expressions for h

$$h = \frac{1}{2}gt^2 \approx \frac{(vt)^2}{2r}$$

$$g \approx \frac{v^2}{r}$$

Note: we have not covered this yet. If this idea (dependence on t^2) is completely new to you, you may want to come back to this part after going through the first chapter on calculus.

The equation $a = v^2/r$ comes even more easily with a little bit of calculus and the use of vectors. See **here**.

Kepler's Third Law

The famous mathematician Johannes Kepler (of whom much more later also), working with observational data from Tycho Brahe, had the following values for the radius

R of the (assumed circular) orbit and the period T (time for completion of one orbit), for five planets.

Orbital data for the six planets known in Kepler's time

	$\overline{r}$ (units of $\overline{r}_{\text{Earth}}$)	T (years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862

On the basis of this data, Kepler published his **third law** (in 1619, about 10 years after the first two). K3 states that

$$T^2 = kR^3$$

The square of the period is proportional to the cube of the radius of the orbit. The data in the table has been scaled so that $k = 1$.

For a circular orbit, the orbital speed, the magnitude of the velocity $v = |\mathbf{v}|$, is constant.

The period times the speed is equal to the circumference.

$$vT = C = 2\pi R$$
$$T = \frac{2\pi R}{v}$$

K3 above says that

$$R^3 = T^2$$
$$= \frac{(2\pi)^2 R^2}{v^2}$$

Hence

$$v^2 \approx \frac{1}{R}$$

We showed above that the acceleration for a circular orbit is

$$a = \frac{v^2}{R} = v^2 \cdot \frac{1}{R}$$

so we conclude that that

$$g = a \approx \frac{1}{R} \cdot \frac{1}{R} = \frac{1}{R^2}$$

if the acceleration of gravity g is directed toward the sun, with a magnitude that is inversely proportional to the square of the distance, then we can explain Kepler's third law by running this chain of reasoning in reverse.

comparing the moon to an apple

Earlier we worked out that the acceleration is

$$a = \frac{v^2}{R}$$

Let's figure out the acceleration of the moon. We make a decision to work in English units for this one.

The moon averages about 237 thousand miles from earth (221.5 - 252.7 thousand miles). The earth's circumference is about 24.9 thousand miles so its radius is about 3.96 miles. Thus, the ratio of the moon's distance to the center of the earth, compared to my distance to the center of the earth, is about 60 : 1 (ranging between 56-64).

What is the moon's velocity? The distance it travels in one complete orbit (in feet) is:

$$2\pi \cdot 2.4 \times 10^5 \cdot 5280$$

The time that takes in seconds is

$$28 \cdot 24 \cdot 3600$$

$$v = \frac{2\pi \cdot 2.4 \times 10^5 \cdot 5280}{28 \cdot 24 \cdot 3600}$$

The acceleration is v^2/R so we square everything except the radius.

$$a = \frac{(2\pi)^2 \cdot 2.4 \times 10^5 \cdot 5280}{(28 \cdot 24 \cdot 3600)^2} = 0.0085$$

That's in feet per second.

We compare this value to the acceleration measured at the surface of the earth, which is 32.2 in the same units. The ratio is 3788, which is just over $(61.5)^2$.

Newton:

> I began to think of gravity extending to the orb of the Moon . . . and computed the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the earth . . . & found them answer pretty nearly. All this was in the two plague years of 1665-1666. For in those days I was in the prime of my age for invention & minded mathematicks and Philosophy more than at any time since.

Part VI

More geometry

Chapter 24

Polygons

Polygons are figures constructed from line segments. They may have 3, 4, 5 or more sides. If the sides are all the same, its a *regular* polygon.

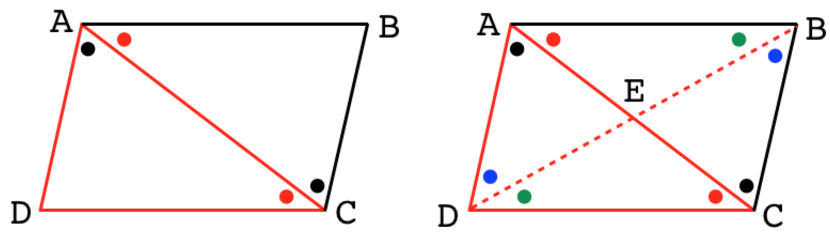
quadrilaterals

Quadrilaterals are four-sided figures. There are a number of types, some of which are:

- square: four sides equal, four right angles
- rectangle: opposite sides equal, four right angles
- rhombus: all sides equal, opposite sides parallel
- parallelogram: opposite sides equal and parallel
- trapezoid: two sides parallel
- kite: adjacent sides equal

parallelogram

Let's look at parallelograms, which can be viewed as two congruent triangles that have been stitched together.



The definition of a parallelogram is that it is a four-sided figure with opposing sides of equal length and parallel. Thus, the interior angles theorem gives us the angle equalities shown.

On the left, we have three angles the same and a shared side, hence $\triangle ABC \cong \triangle ACD$.

Therefore, $AB = DC$ and $AD = BC$.

If we draw the other diagonal and change nothing else, the two triangles are congruent by ASA.

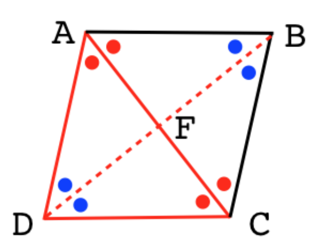
An important property of parallelograms follows: the diagonals cross at their mid-points.

rhombus

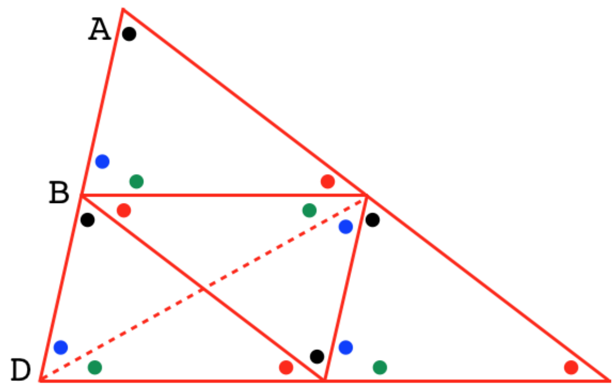
If we further constrain all the sides to be equal, then the half triangles like $\triangle ADC$ become isosceles.

The isosceles triangle theorem says that in a triangle with two sides equal the base angles are equal. The converse is also true.

By the isosceles triangle theorem, all the angles marked with red dots are equal, and all the blue ones as well. Because each quarter triangle has a red and a blue, the central angles are all equal. Therefore, they are all right angles.



Now, finally, let us assemble one whole and two have parallelograms starting with the same figure.

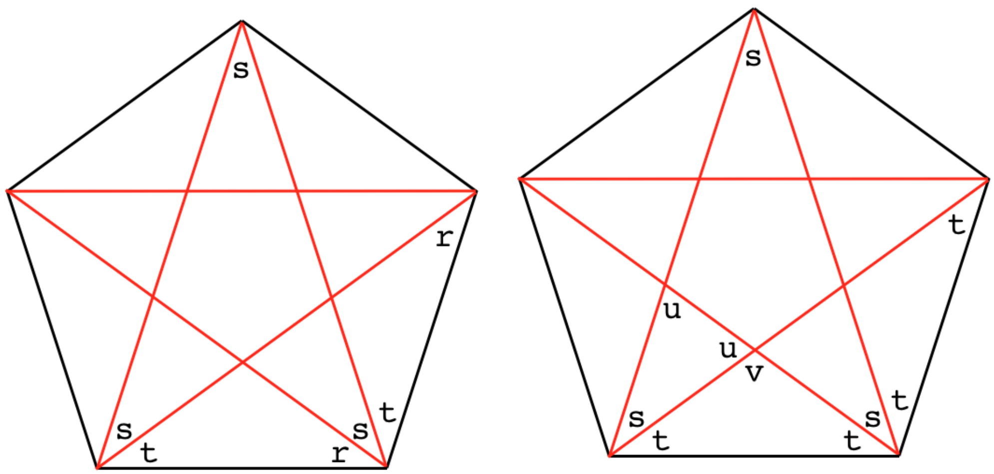


By the properties of the parallelogram, if the adjacent sides AB and BD are equal, all the angles work out and we get four congruent triangles, so the adjacent sides of the large triangle all have equal segments as well.

pentagons

Now we explore some properties of a regular pentagon. The pentagon has five-fold rotational symmetry. Draw all of the internal chords of the figure and label a few angles.

By rotational symmetry each of the five vertices of the pentagon has the same three components, the central one labeled s, and two flanking ones r and t.

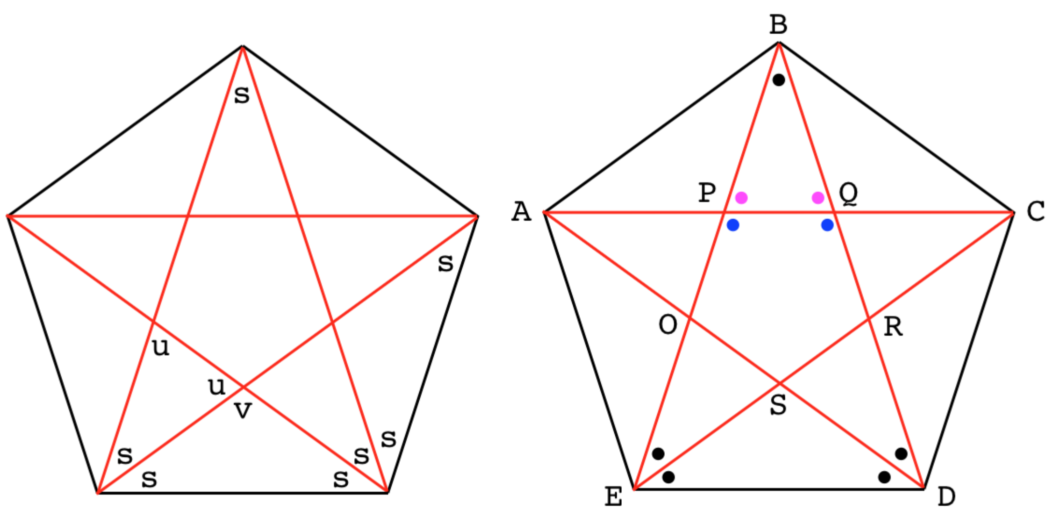


But $r = t$, by Thales' theorem, using two sides of the pentagon. Hence we relabel, immediately (right panel).

We compute two triangle sums:

$$3s + 2t = 4t + s, \qquad 2s = 2t$$

Hence, $s = t$. Relabel t as s (left panel, below):



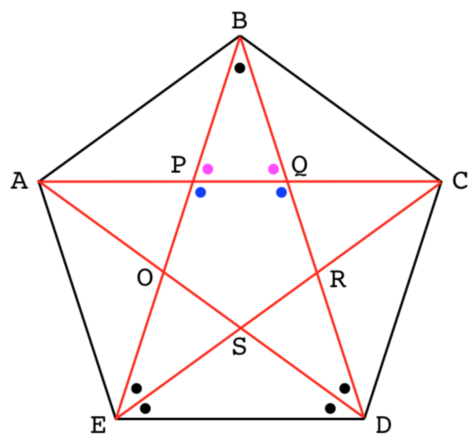
We observe that $5s = \pi$. We again compute two triangle sums:

$$5s = v + 2s, \qquad v = 3s$$

$$5s = 2u + s, \qquad u = 2s$$

Since $v = 3s$, its measure is the same as the vertex angle of the pentagon. Thus the inner figure is also a regular pentagon.

We do not need the angle labels any more, just the equalities.

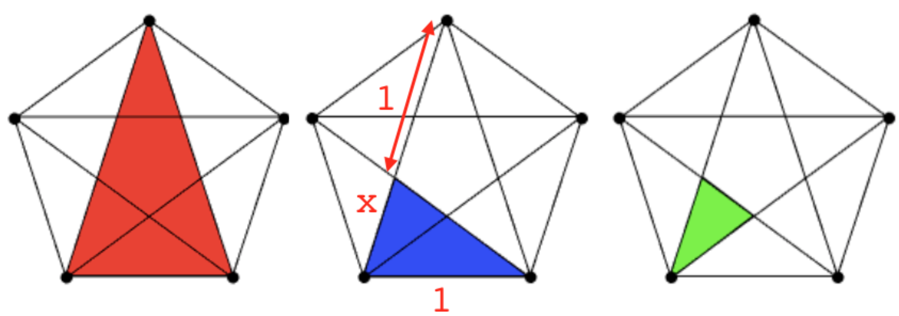


$\triangle BED$ is similar to $\triangle BPQ$. Hence the side of the inner pentagon is in the same measure to the side of the original pentagon as the ratio of PQ to AQ.

Now observe that AC is parallel to ED, because they have the equal alternate interior angles of two parallel lines. (By similar triangles, or simply add the included angles). Our drawing is filled with regular parallelograms, with four sides equal.

One can draw two types of isosceles triangles using the chords and sides of the pentagon. One is tall and skinny, the other short and fat.

The tall, skinny type have base angles equal to s. Here are three examples:



If we take the side length of the original pentagon to be 1, then all the edges of regular parallelograms in the figure also have side length 1, so the long side length of the red triangle is 1 plus some other value, equal to the base of the blue triangle. Let's call that extra part, x.

We use the fact that red and blue are similar and form the ratio ϕ of the long side to the base (red on the left, blue on the right):

$$\frac{1+x}{1} = \frac{1}{x}$$

Rearrange:

$$x^2 + x - 1 = 0$$
$$x = \frac{-1 \pm \sqrt{1+4}}{2}$$

Of course, ϕ is the golden ratio where we have taken the positive branch of the square root:

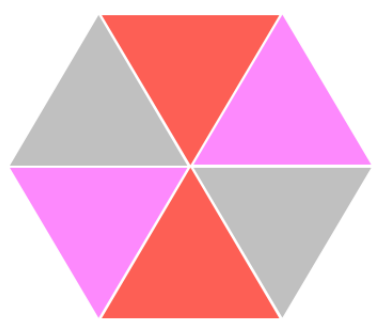
$$\phi = 1 + x = \frac{1+\sqrt{5}}{2}$$

From the similarity of the green triangle, if the side of the inner pentagon is y:

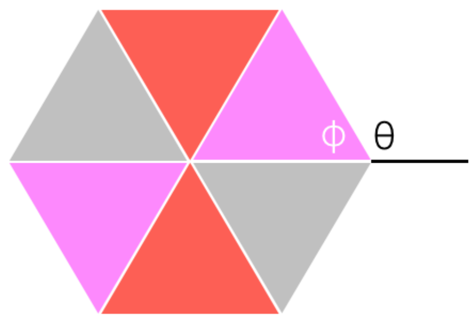
$$\frac{x}{y} = \phi$$
$$y = \frac{x}{1+x}$$

hexagon

A hexagon can be composed of six equilateral triangles. In an equilateral triangle, the three angles are equal and by the angle sum theorem, their measure is $\pi/3$. This matches the requirement for a full $6 \cdot \pi/3 = 2\pi$ angular measure around a point.



We simply note numerical verification of the external angle theorem:



$$\theta = \phi + \phi$$

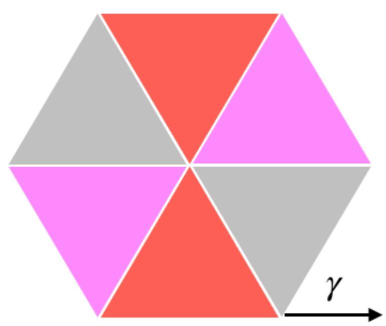
There is a different theorem, why don't we call it the

- Exterior angle sum theorem

to distinguish the angle from the external angle of a triangle, used above.

Imagine walking all the way around a polygon, to do so we must make n turns. What is the measure of the angle at each turn, and what is the total angle turned?

It's tricky, because the diagram we have drawn above is not the right one to use. We need this:



Let's recap what we know for the smaller polygons:

- triangle: $3 \cdot 2\pi/3 = 2\pi$
- rectangle: $4 \cdot \pi/2 = 2\pi$

- pentagon: $5 \cdot 2\pi/5 = 2\pi$
- hexagon: $6 \cdot \pi/3 = 2\pi$

The sum of the external angles is seen to be 2π, which makes sense. We simply turn through $2\pi = 360$ degrees total. Therefore each angle is $2\pi/n$.

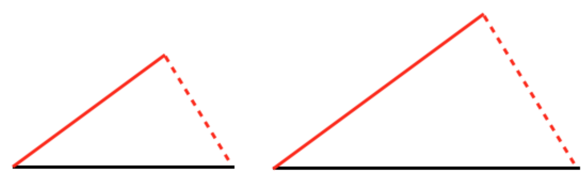
Chapter 25

Similar triangles

To repeat what we said before, some triangles may be *similar* but not congruent.

Similarity means that the three angles are the same but the triangles are of different overall sizes. We might say that they are the same but *scaled* differently.

We can call this AAA (angle-angle-angle).



Actually, there are three statements which are equivalent, but need to be proved equivalent as theorems. This isn't always done carefully.

- Two triangles are similar if they have the same three angles.

- Two triangles are similar if their sides are parallel to one another.

- Two triangles are similar if their sides are in the same ratio to each one another.

Because of the angle sum theorem, if any two angles of a pair of triangles are known to be equal, then the third one must be equal as well.

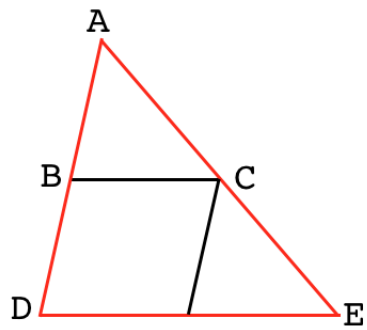
- Two triangles are similar if they have two angles known to be equal.

For similar triangles, the three corresponding pairs of sides are in the same proportions, but re-scaled by a constant of proportion. This is known as the AAA similarity theorem.

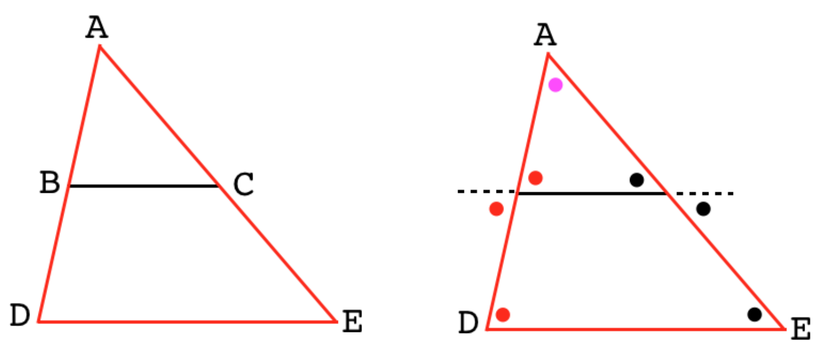
Before getting there we introduce another theorem.

midline theorem

Theorem: The line segment BC connecting the midpoints of two sides of a triangle is parallel to the third side, and is congruent to half of it.

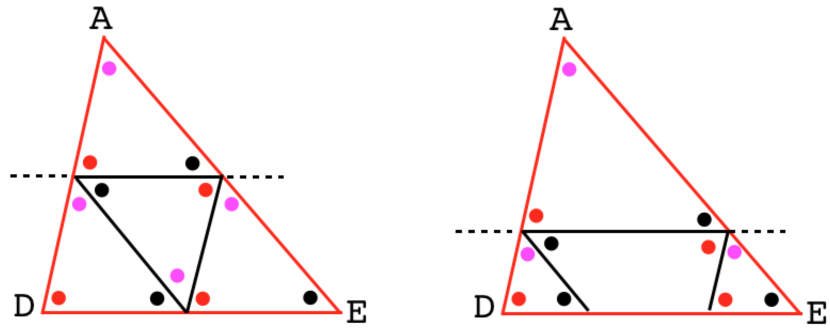


Earlier we said that in a triangle like the one in the figure below, if we are given that BC is parallel to DE, then by the a combination of the alternate interior angles theorem and the vertical angles theorem, we can show that $\triangle ABC$ and $\triangle ADE$ have three angles the same.

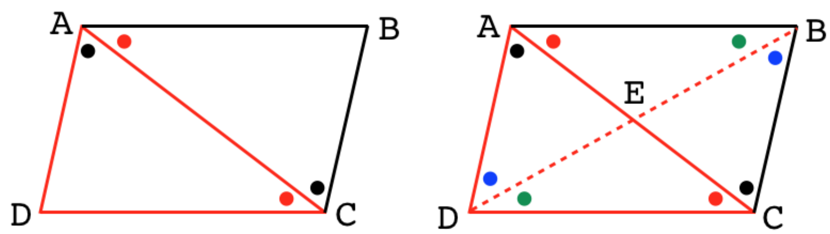


Their sides are also in proportion. How can we prove that?

Let us start with the case where $AB = BD$. Draw similar line segments parallel to AE and parallel to AD.



Notice that something nice happens on the left, in the case where $AB = BD$. To see why, it may be useful to take a side track and look at parallelograms, which are basically triangles that have been stitched together.



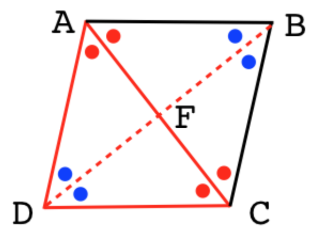
The definition of a parallelogram is that it is a four-sided figure with opposing sides parallel. Thus, the interior angles theorem gives us the angle equalities shown. On the left, we have three angles the same and a shared side, hence $\triangle ABC \cong \triangle ACD$. We have shown that $AB = DC$ and $AD = BC$.

If we draw the other diagonal and change nothing else, opposing triangles are congruent by ASA. Therefore, the diagonals cross at their midpoints.

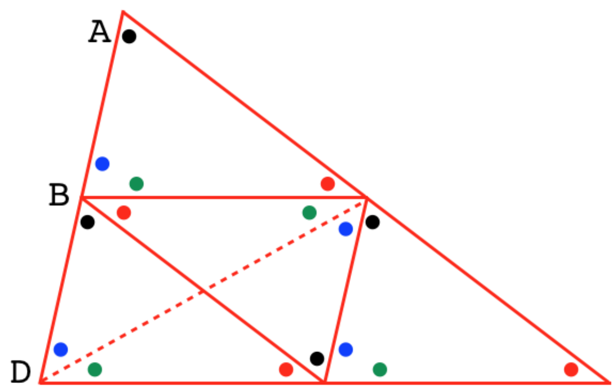
If we further constrain all the sides to be equal, then the half triangles like $\triangle ADC$ become isosceles.

The isosceles triangle theorem says that in a triangle with two sides equal the base angles are equal. The converse is also true. Note that in the proof of the isosceles triangle theorem, we do not use any facts about similar triangles, only congruent ones!

By the isosceles triangle theorem, all the angles marked with red dots are equal, and all the blue ones as well. Because each quarter triangle has a red and a blue, the central angles are all equal. Therefore, they are all right angles.



Now, finally, let us assemble one whole and two half parallelograms starting with the same figure.



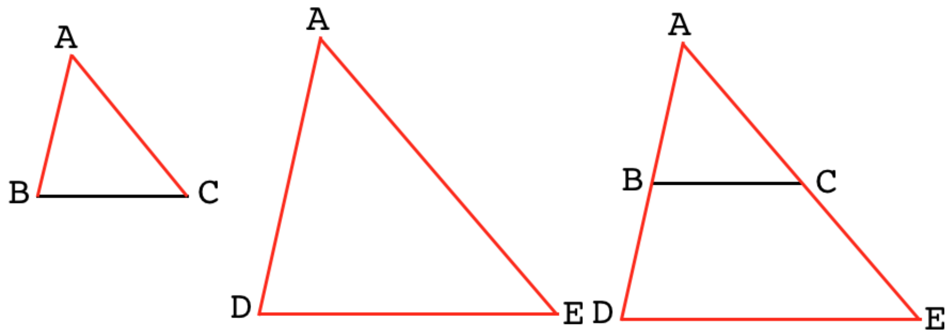
By the properties of the parallelogram, if the adjacent sides AB and BD are equal, all the angles work out and we get four congruent triangles, so the adjacent sides of the large triangle all have equal segments as well.

So finally, what about the case where the adjacent segments are not equal? Good question!

AAA similarity theorem

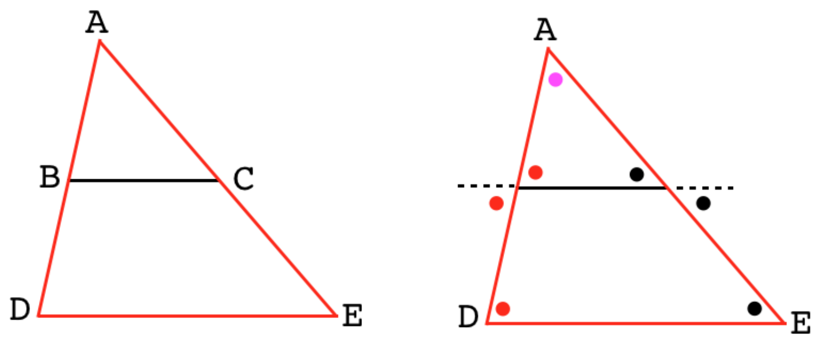
The proof comes in two parts, first for the angles, and then for the ratio of sides.

The main theorem used for angle identity is the alternate interior angles theorem, which is an if and only if theorem: such angles are equal $\iff$ there is a transversal of two parallel lines.

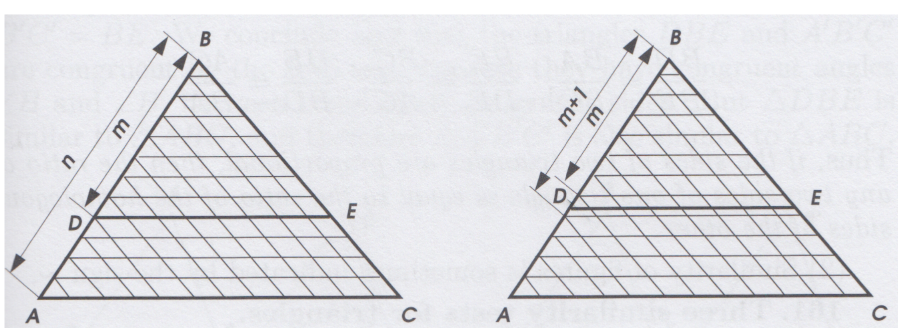


Suppose we have $\triangle ABC$ and separately $\triangle ADE$ and we are given that the angle at vertex A is the same for both, and when we draw $\triangle ABC$ on top of $\triangle ADE$, BC is parallel to DE.

By the alternate interior angles theorem, $\angle ABC = \angle ADE$ and $\angle ACB = \angle AED$.



Proving that the ratios of sides are all the same is a bit harder and often not even done explicitly. I found a good proof in Kiselev.



(Note: he draws $\triangle BDE$ smaller than $\triangle BAC$).

There are two cases. The first is when the lengths of BA and BD are commensurable. That is, there is some unit length such that for integers m and n, $BD = m$ and $BA = n$.

Divide the side as shown. Draw lines parallel to BC and AC.

Then BE and BC will be divided into congruent parts, m and n for each, respectively. The same thing happens on the bottom. It is clear that

$$\frac{m}{n} = \frac{BD}{BA} = \frac{DE}{AC} = \frac{BE}{BC}$$

The second case is shown in the right panel above. BD and BA are not commensurate and there is some small remainder when dividing m into n. But

$$\frac{m}{n} \approx \frac{BD}{BA}, \qquad \frac{m}{n} \approx \frac{DE}{AC}, \qquad \frac{m}{n} \approx \frac{BE}{BC}$$

And furthermore, by choosing n larger, we can make the remainder as small as we please. Thus we obtain, in the limit and n gets very large:

$$\frac{m}{n} = \frac{BD}{BA} = \frac{DE}{AC} = \frac{BE}{BC}$$

for the second case as well.

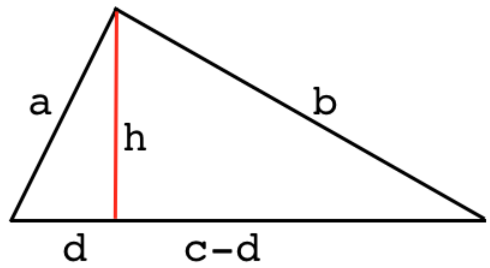
another idea

Euclid doesn't get around to proving the AAA similarity theorem until Book VI.

If you look at the chapter on the Pythagorean theorem, you will see that Euclid's proof from $PI.47$ depends only on triangle congruence and not anything about similarity. Therefore, we can rely on it here.

Here is a sketch for a proof.

We also have an algebraic proof of Pythagoras which starts from this figure:



and assumes the side ratios of similar right triangles

hypotenuse to short side

$$\frac{a}{d} = \frac{b}{h} = \frac{c}{a}$$

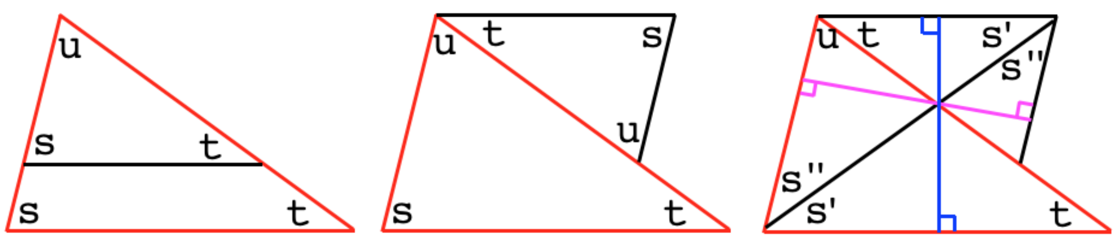
hypotenuse to long side

$$\frac{a}{h} = \frac{b}{c-d} = \frac{c}{b}$$

and so goes on to prove the Pythagorean theorem. You can run this proof backwards from the theorem to the side ratios result. Therefore, AAA similarity is established for right triangles.

Then, any triangle can be decomposed into two right triangles. So, combining the results for the two sub-triangles, we would have the result for the general case.

Here is a sketch of that part:



We have two triangles similar because the angles are the same (left panel). Flip the smaller triangle up and form a partial parallelogram. The original small triangle and the flipped version are congruent by SAS.

From the parallel sides we get the angle equalities in the right panel. Draw the two altitudes.

Now we have various similar right triangles. Composing the results for the right triangles, we can build up to the main result.

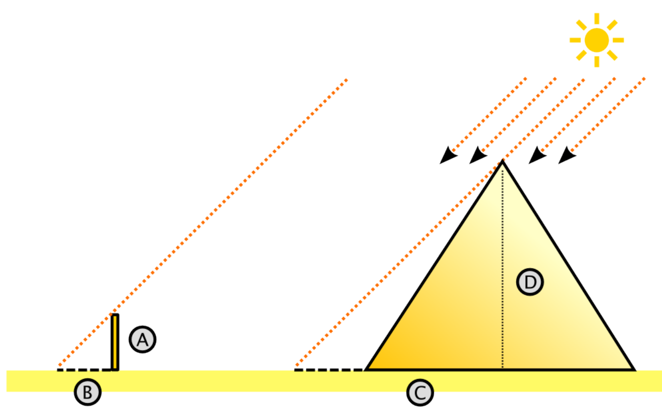
pyramid height

As we said, Thales was from Miletus and he lived around 600 BC. Thales is believed to have traveled extensively and was likely of Phoenician heritage. As you probably know, the Phoenicians were famous sailors who founded many settlements around the Mediterranean.

They competed with the mainland Greeks and later with the Romans for colonies, and their major city, Carthage, was destroyed much later by the Romans in the third Punic War.

During his travels, Thales went to Egypt, home to the great pyramids at Giza, which were already ancient then. They had been built just around around 2560 BC (dated by reference to Egyptian kings) and were already 2000 years old at that time!

The story is that Thales asked the Egyptian priests about the height of the Great Pyramid of Cheops, and they would not tell him. So he set about measuring it himself. The current height is 480 feet. He used similar triangles.

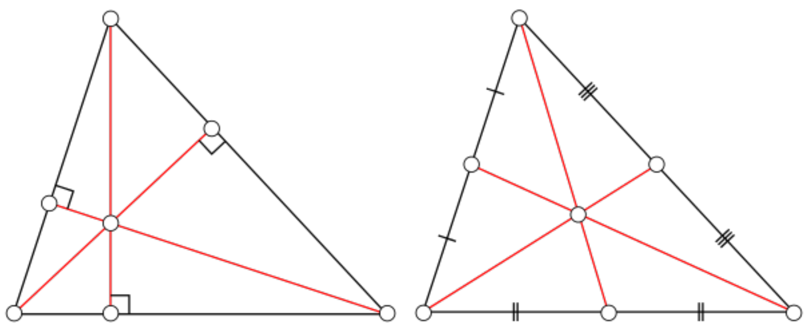


Chapter 26

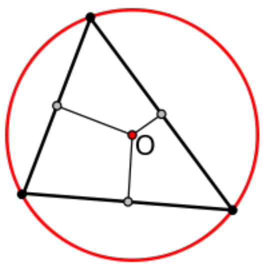
Special points

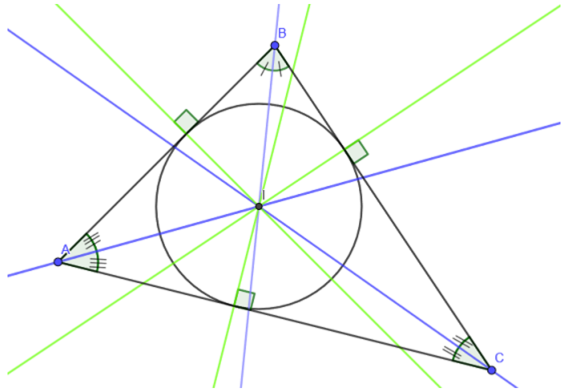
Special points in triangles include the:

- orthocenter: where altitudes cross
- centroid: where the medians (lines to the midpoints of sides) cross

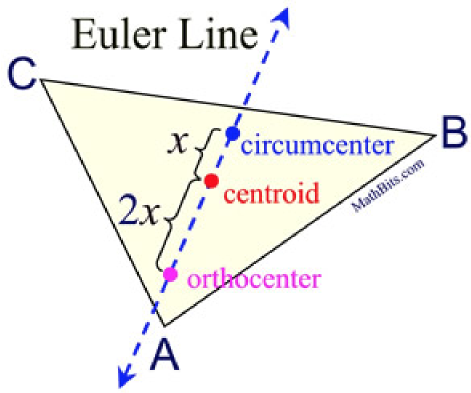


- circumcenter: the center of the circle where all three vertices lie, also, where perpendiculars to side midpoints cross
- incenter: where angle bisectors cross





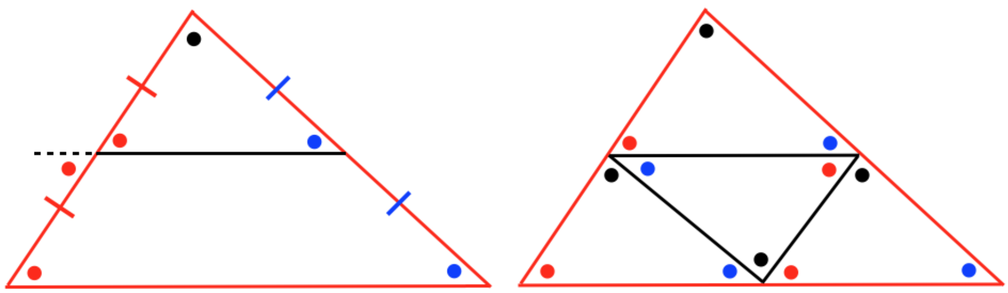
Euler showed that the first three of these points lie on one line.



centroid

We'll start with the centroid. The centroid is where bisectors of opposing sides cross.

Consider the triangle in the figure below (left panel).



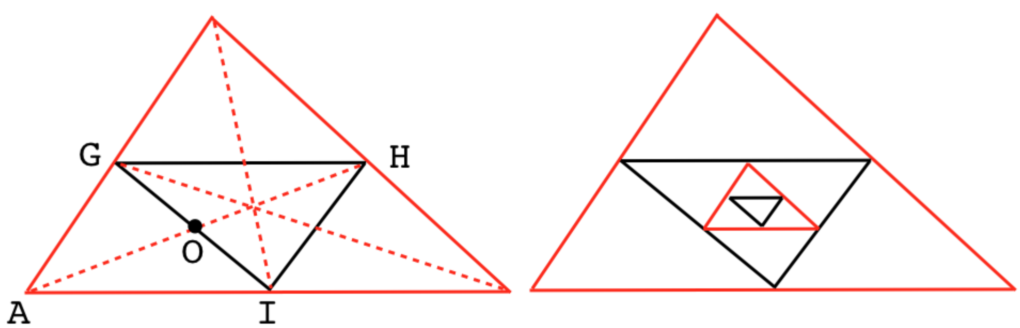
Draw a line segment parallel to the base and connecting to the midpoint of the left side. Then, by the alternate interior angle theorem and the vertical angle theorem, the two angles marked with red dots in the middle are equal to the red dotted angle at the base.

Therefore, by three angles the same, the small upper triangle is similar to the large one. The ratio of similar sides is 1 : 2.

But this can be done on the right side as well, and then the same for all three vertices of the original triangle (right panel).

By the triangle sum theorem and also by the alternate interior angle theorem, the angles in the interior triangle are equal to other angles as indicated. By shared sides, the four small triangles are congruent.

Now draw lines from each vertex to the midpoint of the opposing side. $GHIA$ is a parallelogram, by the angle equalities just proven. The two diagonals of a parallelogram cross at their midpoints. Therefore O is the midpoint of the side GI and the same line that connects A to midpoint H also connects H to midpoint O.



Therefore the centroid of $\triangle GHI$ is also the centroid of the parent. This process can be repeated as many times as we please (right panel). The triangles get smaller and eventually tend to a point. That point is on all three midpoint segments. Therefore, the centroid is a single point.

$\square$

Note: this is actually a special case of Ceva's theorem, which we proved previously. But I really like the proof above, which is from Lockhart.

If you know about vectors, it is a good exercise to prove Ceva's theorem for the centroid using vectors.

algebra of the centroid

We can locate the centroid by imagining that we find successive midpoints of a length from opposite ends left and right.

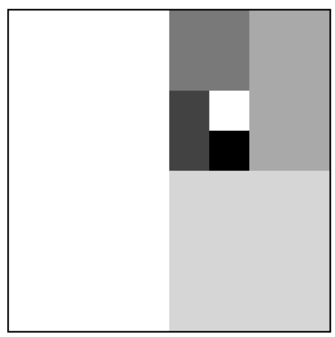
The first point is at 1/2 of the length (point O on $\triangle GHI$), the second comes back from vertex H by 1/4 so is at 0.75 (on the right edge of the small red triangle in the right panel, above). The third is at $0.5 + 1/8$ (on the left edge of the smallest black triangle).

Every second round we get closer to the centroid by advancing from the left by

$$S = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \dots$$

Now, we can either assume this sum is finite (for now) or recognize that it is certainly smaller than

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1$$



So if

$$S = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \dots$$

then

$$2S = 1 + \frac{1}{4} + \frac{1}{16} + \dots$$

and

$$3S = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots$$
$$= 1 + 1$$
$$S = \frac{2}{3}$$

□

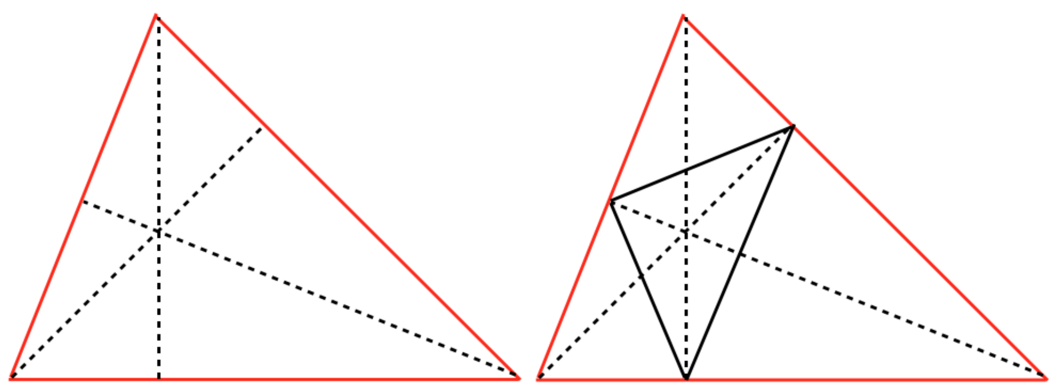
Chapter 27

Orthocenter

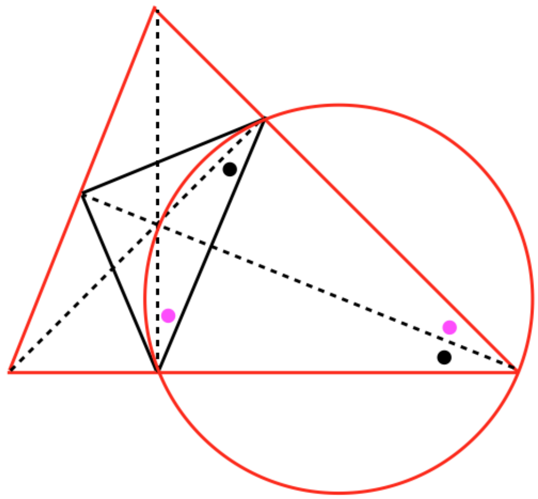
orthocenter

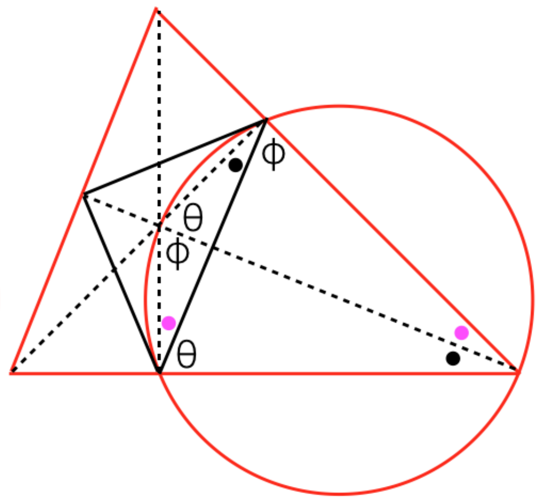
An altitude of a triangle is a line extended from a vertex so as to form a right angle with the opposing side. The *orthocenter* is the point where the three altitudes of a triangle meet.

Assume for now that the three altitudes *do* meet at a single point, we will come back to this question later.



Above we have drawn the altitudes (left panel) and connected the points where the altitudes meet the sides. We can prove something interesting about the inset triangle. We will prove that the incenter of that triangle is equal to the orthocenter of the bigger one.



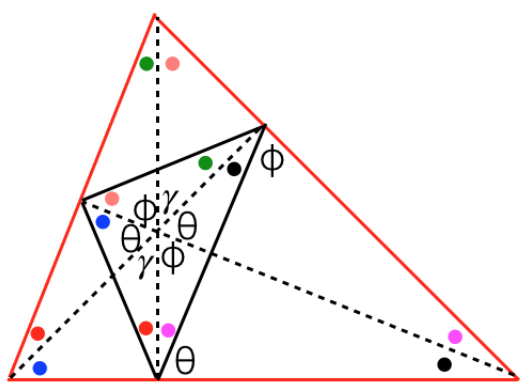


The dotted lines form right angles at the sides. Therefore, we can use the common dotted line as a diameter and draw a circle that includes all four points.

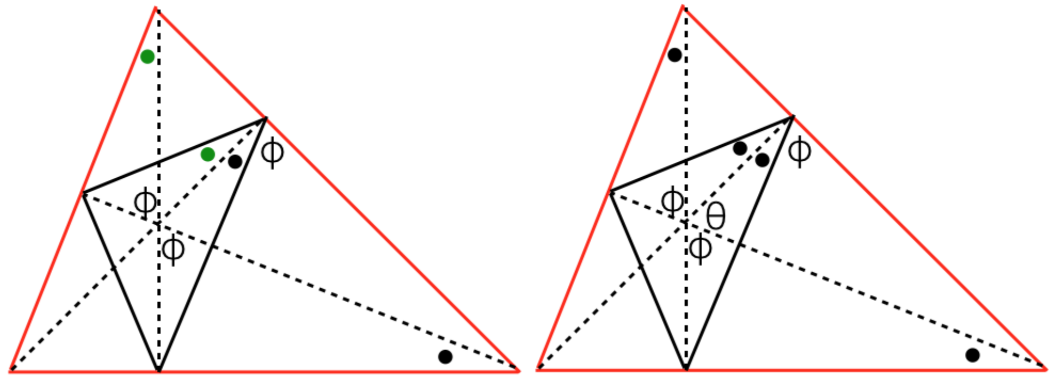
Despite what it looks like, the figure is not symmetrical across the diameter. For example, the angles marked with black dots and those marked with magenta dots are not equal.

However, now we can use our theorems about arcs subtended by an angle on the perimeter of the circle. The two angles marked with black dots subtend the same arc. Similarly for magenta. The same is also true for the pairs of angles marked ϕ and θ.

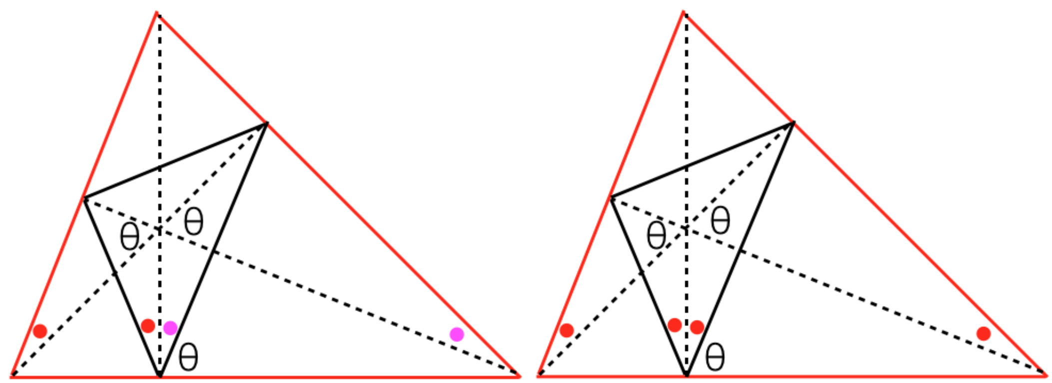
We use symmetry to fill in some more equalities.



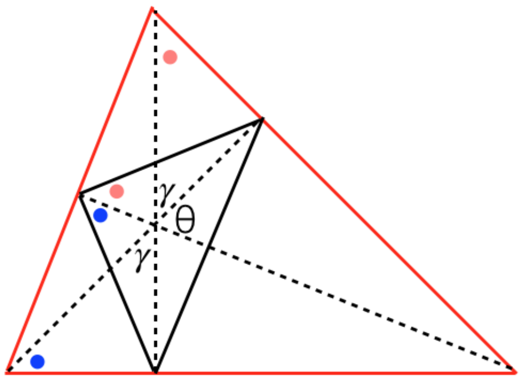
This looks like a mess. But let us look more carefully at one set of angles, ϕ and the black and green.

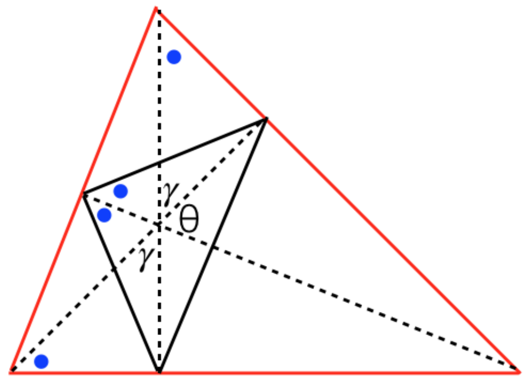


Notice that, at the bottom, ϕ is complementary to black, but the top left, ϕ is complementary to green. Hence black and green must be equal (right panel). We carry out the same exercise for θ, red and pink.

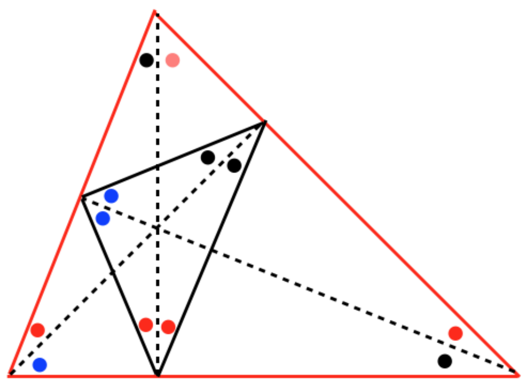


And then, for γ, blue and salmon.





Restoring dots



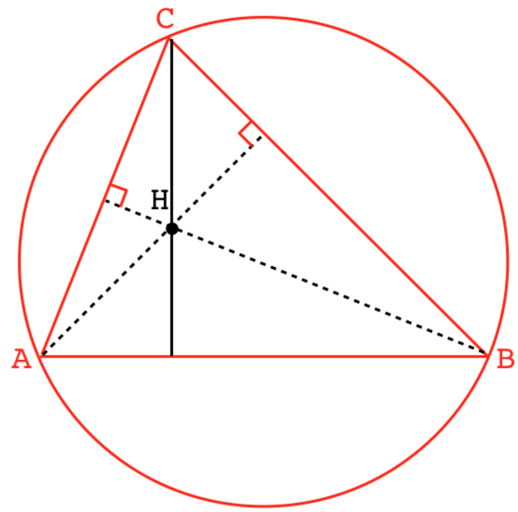
We can see that the dotted lines are angle bisectors for the small triangle.

Thus, the orthocenter of the large triangle and the incenter of the triangle inscribed between the points along the base, are the same point.

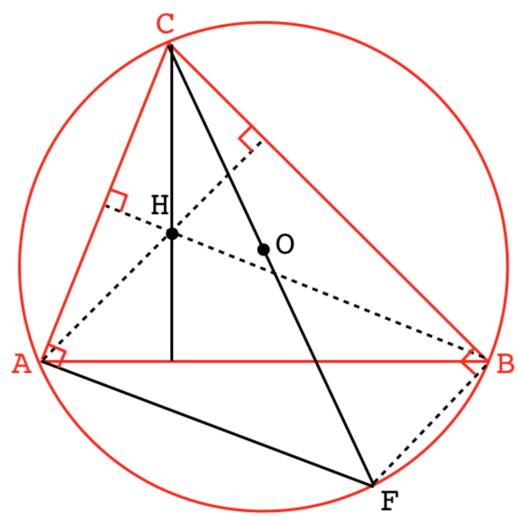
□

Orthocenter exists

Because we can (and have already) extended Ceva's general proof to the specific case of the orthocenter (altitudes meeting at a point), we don't need a specific proof. Nevertheless, it's good practice. Here is a nice short one.



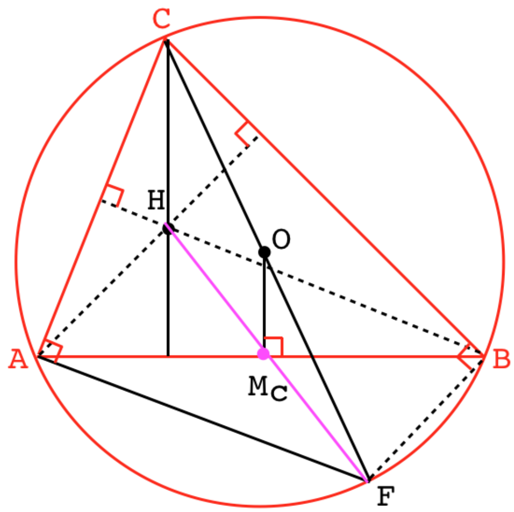
We have $\triangle ABC$, with two altitudes drawn (and marked as right angles where they hit the sides), and a third line from vertex C that we need to show makes a right angle with side AB. We've also drawn the circumcircle of the triangle.



Draw the diameter of the circle, and form right angles at vertexes A and B by inscribing a right triangle as each.

Since AH is on a line that forms a right angle with BC and so is BF, we have that AH is parallel to BF. Similarly, BH is parallel to AF, so $AHBF$ is a parallelogram.

We draw the other diameter of the parallelogram.



Diameters of a parallelogram cross at their midpoints. Therefore, $AM_c = M_cB$ and so the line from the center to M_c is a perpendicular bisector of AB. As well, $HM_c = M_cF$.

Because $CO = OF$ and $HM_c = M_cF$, OM_c is a *midline* of $\triangle CHF$

Therefore, $\triangle CHF$ is similar to $\triangle OM_cF$. As a result, CH is parallel to OM_c.

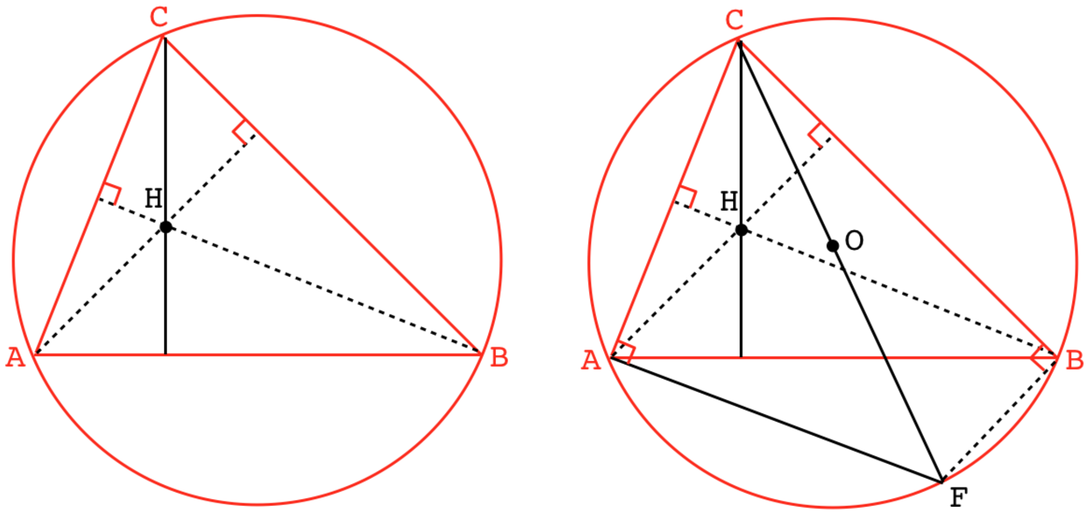
Therefore its extension forms a right angle with AB, just as OM_c does.

$\square$

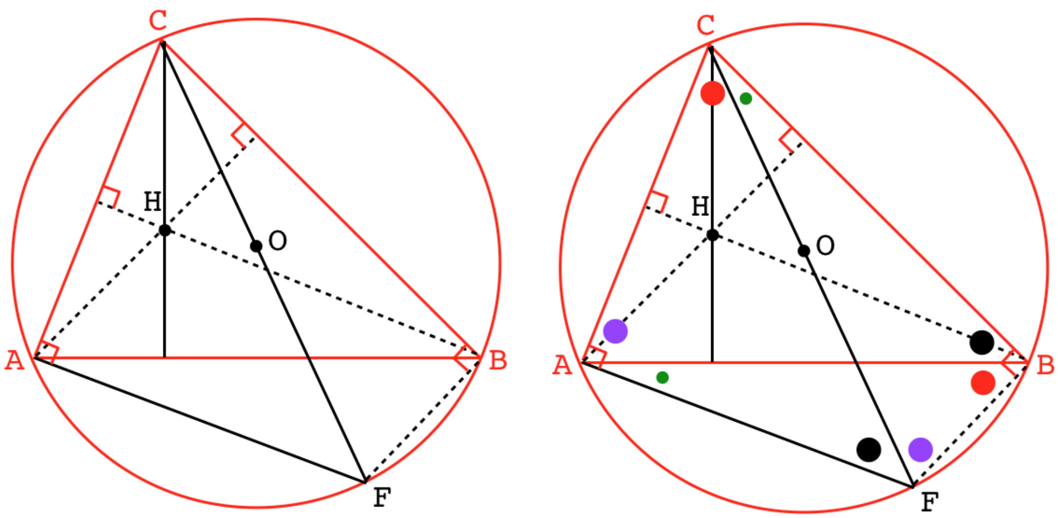
I played with this a lot, looking for a proof using the many equalities in the figure. Remember that we must never assume at any point that the altitude from C forms a right angle with AB. Whenever I thought I had it, I discovered I had violated that rule.

We'll refer to the altitudes when needed as h_a and h_b which intersect the sides opposite to vertices A and B, respectively.

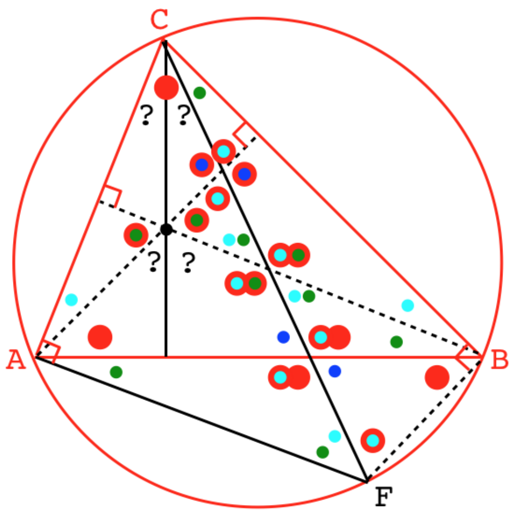
Draw the diameter, the right angles at vertices A and B, and draw the parallelogram. (Each bit of text references the figure below, which are before and after shots).



Find the various equal angles that subtend equal arcs. Only those on the perimeter are allowed!



Here is where I stopped working on it. We need one of those question marks, by itself.

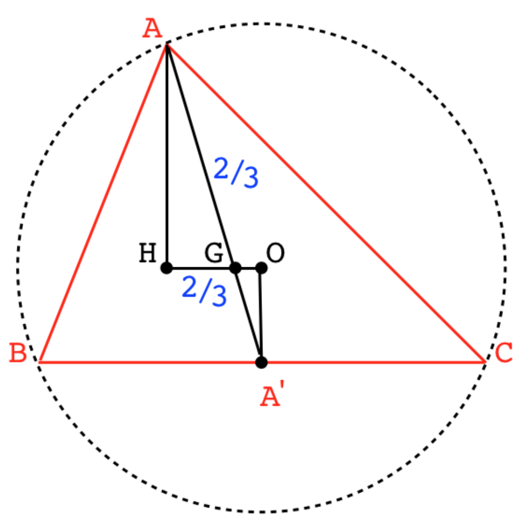


Euler's proof

Below we give an alternative proof, due to Euler, which is stunning, following

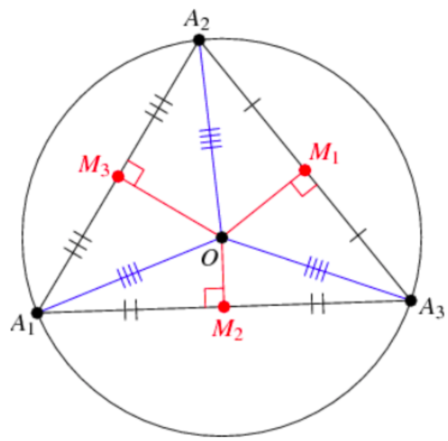
https://artofproblemsolving.com/wiki/index.php/Orthocenter

A copy of their figure:



The orientation is reversed from what we had above. First, the point O is the circumcenter of the triangle: the center of the circle which contains all three vertices of the triangle.

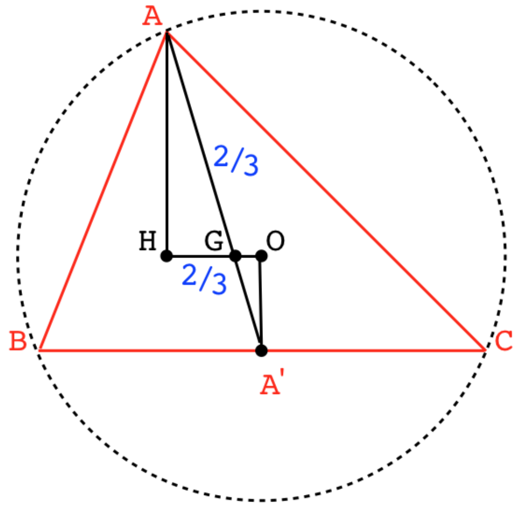
Clearly, this circle has a center. The classic construction is to bisect each side (here BC is bisected at A'), and erect a perpendicular. The point where the three perpendiculars cross is the circumcenter, which is the center of the circle.



So, assume we have done this and that point is O.

The next point, G, is the centroid. One way to find this point is to draw all three lines connecting vertices with the midpoints of the opposite side (AA'). However, if you recall, the distance from the vertex A to G is twice the distance from the midpoint A' to G. Hence we draw point G using arithmetic.

Now, extend OG by twice its length, to H. ($2OG = GH$).



Because AG is twice $A'G$ and GH is twice OG and the two triangles share both angle $\angle OGA'$ (equal to $\angle AGH$), they are similar triangles.

Since $\angle A'OG$ is a right angle, therefore so is $\angle AHG$. This means that AH is perpendicular to BC. Thus, AH is a part of the altitude from A to BC (the whole altitude is not shown).

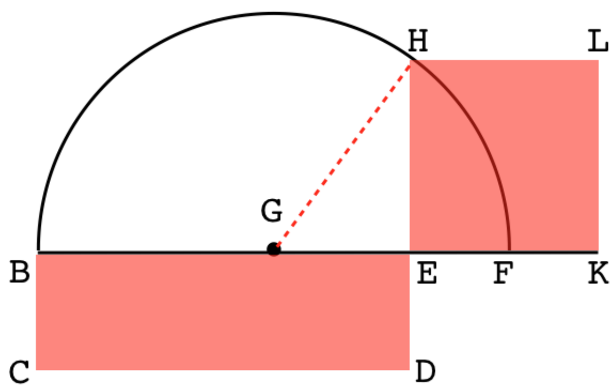
The same construction could be done for the other two vertices, each time ending at H. This shows that H is unique, and that H is on all three altitudes.

This proof also demonstrates that the orthocenter, centroid and circumcenter lie on a single line, and that the distance from centroid to orthocenter is twice that from centroid to circumcenter.

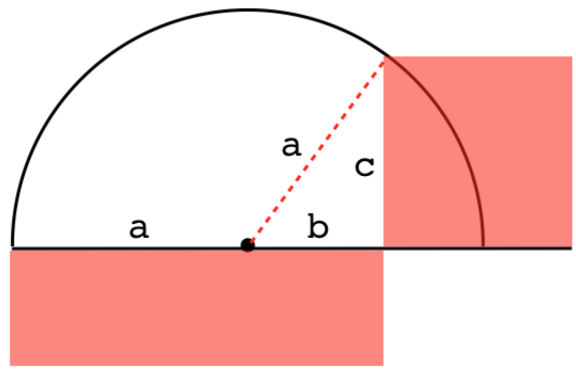
Chapter 28

Hippocrates

Hippocrates of Chios (470-410 BC) was a major figure in Greek geometry. (Not to be confused with the physician of the same name, from Kos). Hippocrates focused on quadrature, the process of constructing (with straight-edge and compass) a square with area equal to a given geometric figure, particularly curved figures, called lunes. Here is one of the first of these—construction of the square equivalent to a given rectangle.



The construction says to: (i) extend BE horizontally, (ii) mark off the same distance as DE to construct EF, (iii) find the midpoint G of BF, (iv) draw the half-circle of radius BG, (v) extend DE up to meet the circle at H, construct the square of side the same length as EH.



As suggested by the dotted line in the figure, the proof invokes the Pythagorean theorem. The long side of the rectangle is $a + b$, while its short side is $a - b$, so the area is

$$A = (a + b)(a - b) = a^2 - b^2$$

but Pythagoras says that is equal to c^2.

$\square$

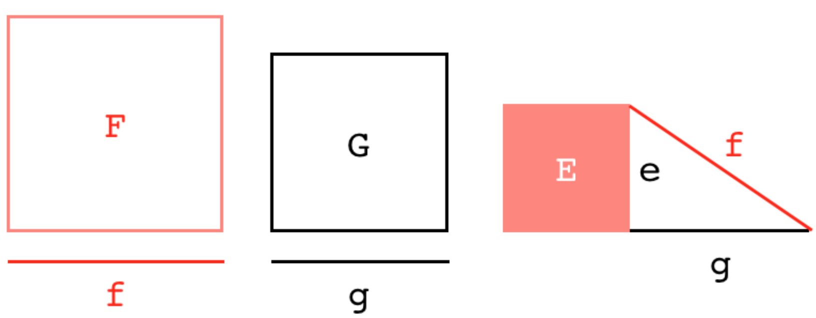
This is a slight restatement of our proof about the geometric mean.

The side of the square, c is the geometric mean of the sides of the rectangle.

$$c = \sqrt{(a + b)(a - b)}$$

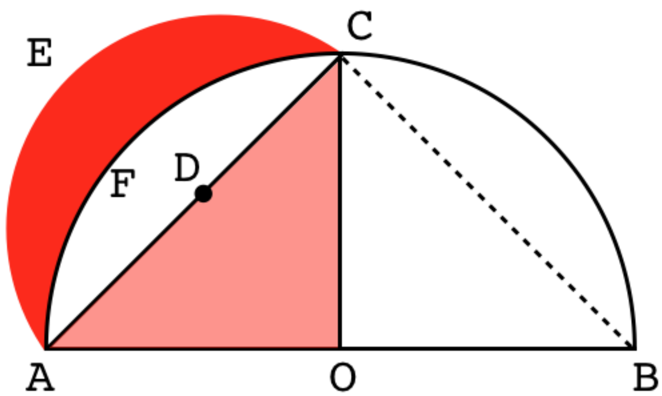
other constructions

Hippocrates "squared" rectangles, triangles and polygons. A lot of his constructions depended on Pythagoras as suggested by this figure:



where two squares resulting from manipulation of part of a polygon need to be subtracted to obtain the final result.

Hippocrates moved on to curves, trying to find squares with area equal to that under or between two curves. That turns out to be a class of problems where few have solutions (in fact, only five, according to Dunham). Famously, it is impossible to square the circle. However, here is one that is possible, it is an example of (the) quadrature of the lune.



We will prove that the two shaded regions are equal in area.

Consider the smaller semicircle with base ADC, which is also the hypotenuse of the right triangle. Let radius AD be equal to r. Let the large semicircle have radius AO equal to R. Pythagoras tells us that

$$R^2 + R^2 = (2r)^2$$

$$R^2 = 2r^2$$

Let the area of the triangle be T. (Its value is $R^2/2$ but that's not needed).

The segment of the larger semicircle (white) is the area of the quadrant minus the area of the triangle

$$\pi \frac{R^2}{4} - T$$

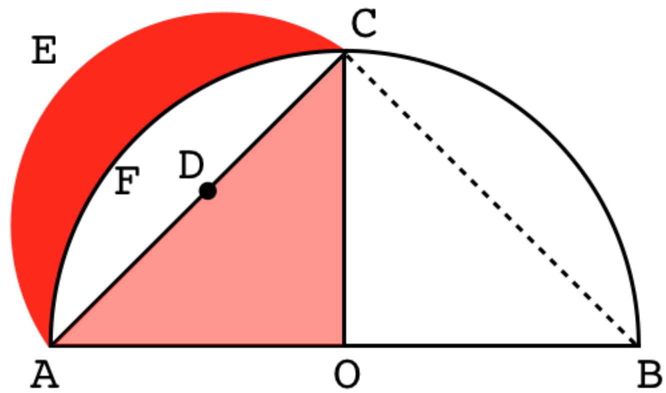
The area of the red lune is the area of the small semicircle minus the white area

$$\pi \frac{r^2}{2} - \ [\ \pi \frac{R^2}{4} - T\ ]$$

$$= \pi \frac{r^2}{2} - \pi \frac{2r^2}{4} + T$$

$$= T$$

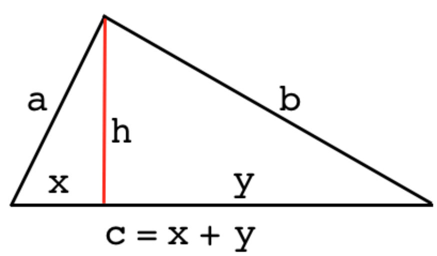
Which is just the area of the triangle.



Chapter 29

Heron's formula

Heron's Formula can be used to compute the area of a triangle from the lengths of its sides. It is a simple formula that does not explicitly include the altitude h or the parts of side c.



If s is the semi-perimeter

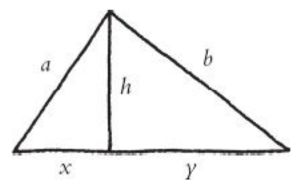
$$s = \frac{1}{2}(a + b + c)$$

then

$$A = \sqrt{s + (s - a) + (s - b) + (s - c)}$$

Proof.

From Lockhart.



Side c is split into x and y. We can write three equations:

$$x^2 + h^2 = a^2$$
$$y^2 + h^2 = b^2$$
$$x + y = c$$

Our objective is an equation that contains only a, b and c. Lockhart gives us a target for the first part of the derivation:

$$2xc = c^2 + a^2 - b^2$$

Let's just start manipulating equations to get there. Subtract the second from the first:

$$x^2 - y^2 = a^2 - b^2$$

Square the third

$$x^2 + 2xy + y^2 = c^2$$

Add the two new equations

$$2x^2 + 2xy = c^2 + a^2 - b^2$$

Substitute for y

$$2x^2 + 2x(c - x) = c^2 + a^2 - b^2$$
$$2xc = c^2 + a^2 - b^2$$

Finally a slight rearrangement:

$$x = \frac{c^2 + a^2 - b^2}{2c} = \frac{c}{2} + \frac{a^2 - b^2}{2c}$$

This says that to find the point where c is divided into x and y, we move from the center $c/2$ a distance of $(a^2 - b^2)/2c$.

The corresponding equation for y is

$$y = \frac{c}{2} - \frac{a^2 - b^2}{2c}$$

which is easily checked by adding together the final two equations, obtaining $x+y = c$.

For the area, we will need h somehow. It is easier to use h^2.

$$h^2 = a^2 - x^2$$
$$= a^2 - \frac{(c^2 + a^2 - b^2)^2}{(2c)^2}$$

The area squared is

$$A^2 = \frac{1}{4}c^2h^2$$
$$= \frac{1}{4}c^2a^2 - \frac{1}{4}c^2\frac{(c^2 + a^2 - b^2)^2}{(2c)^2}$$

Lockhart:

> the algebraic form of this measurement is aesthetically unacceptable. First of all, it is not symmetrical; second, it's hideous. I simply refuse to believe that something as natural as the area of a triangle should depend on the sides in such an absurd way. It must be possible to rewrite this ridiculous expression...

Here's a start:

$$16A^2 = (2ac)^2 - (c^2 + a^2 - b^2)^2$$

This is much better. We notice that it is a difference of squares. First

$$16A^2 = \ [\ 2ac + (c^2 + a^2 - b^2)\ ]\ [\ 2ac - (c^2 + a^2 - b^2)\ ]$$

But that has within it two squares, namely $(a+c)^2$ in the first term on the right-hand side, and $(a - c)^2$ in the second.

$$= \ [\ (a + c)^2 - b^2)\ ]\ [\ b^2 - (a - c)^2\ ]$$

$$= (a+c+b)(a+c-b)(b+a-c)(b-a+c)$$

So

$$A = \sqrt{\frac{a+b+c}{2} \cdot \frac{a+c-b}{2} \cdot \frac{a+b-c}{2} \cdot \frac{-a+b+c}{2}}$$

At this point, we recognize the semi-perimeter $s = (a+b+c)/2$ and then we see that each of the other terms is s minus one of the sides. For example:

$$\frac{a+c+b}{2} - b = \frac{a+c+b-2b}{2} = \frac{a+c-b}{2}$$

So we obtain

$$A = \sqrt{s \cdot (s-a) \cdot (s-b) \cdot (s-c)}$$

check

As a simple example, if we have a right triangle with sides 3,4,5, then the area is one-half of 3 times 4 = 6. The semi-perimeter is s

$$s = \frac{(3+4+5)}{2} = \frac{12}{2} = 6$$

We have

$$A = \sqrt{6(6-5)(6-4)(6-3)} = \sqrt{6(1)(2)(3)} = 6$$

Part VII

The real numbers

Chapter 30

Rational numbers

The integers are great, they give us an infinite supply of numbers.

However, there is a problem with division. For

$$p \in \mathbb{N}, \quad q \in \mathbb{Z}$$

very often the result of $p \div q$ is not contained in $\mathbb{N}$ or even in $\mathbb{Z}$ — the result is not an integer. We say these sets are not *closed* under division.

For example $3 \div 2 = ?$

So, we just leave the result as

$$\frac{p}{q} = \frac{3}{2}$$

where p/q is in "lowest terms", i.e. they have no common factor other than 1. Of course if p and q have a common divisor, then we can divide both top and bottom by the largest common divisor.

q must not be zero because division by zero is not defined.

theorem

• The arithmetic combinations of two rational numbers are rational numbers.

Proof.

This is just basic algebra:

$$\frac{p}{q} + \frac{r}{s} = \frac{ps + rq}{qs}$$
$$\frac{p}{q} - \frac{r}{s} = \frac{ps - rq}{qs}$$
$$\frac{p}{q} \cdot \frac{r}{s} = \frac{pr}{qs}$$
$$\frac{p}{q} \div \frac{r}{s} = \frac{p}{q} \cdot \frac{s}{r} = \frac{ps}{qr}$$

theorem

• Between *any* two rational numbers it is always possible to find another rational number.

Consider two rational numbers, not equal. Let

$$s = \frac{p}{q} \quad t = \frac{p'}{q'}$$

Suppose $s < t$.

The *average* of these two numbers is:

$$r = \frac{1}{2} [\ s + t\ ]$$

Then

$$2r = s + t$$
$$2r - 2s = t - s$$

We have that $s < t$, so $t - s > 0$ and then

$$r - s > 0$$
$$r > s$$

A similar argument will show that

$$r < t$$

so

$$s < r < t$$

□

Thus, one can always find a new rational number that lies between two known rational numbers. In particular, there is no *smallest* positive rational number.

decimal representation

Every rational number can be represented as a decimal, using the method called long division.

Consider 1/2

$$2\overline{)1.000}$$

We say that 2 does not *go into* 1, since $2 > 1$, so we have the first part of our result as 0, followed by a decimal point. But 2 does go into 10 exactly 5 times, giving 0.5. The remainder is zero and so the division process terminates.

Consider 1/8.

$$8\overline{)1.000}$$

∘ 8 goes into 10 once, leaving 2 as remainder

∘ 8 goes into 20 twice, leaving 4.

∘ 8 goes into 40 exactly 5 times with no remainder.

The result is 0.125.

The other possibility is that in going through the process a remainder comes up that has been seen previously. If this happens then the sequence will repeat forever.

If we don't terminate with zero, then this must eventually happen, because there are only as many as q possible remainders.

Thus, for example

$$1/7 = 0.142857142857\ldots$$

which contains 142857, repeating.

decimals to fractions

Conversely, every repeating decimal can be represented as a rational number. For example

```
      1 x r =      0.142857142857...
1000000 x r = 142857.142857...
 999999 x r = 142857
```

$$r = \frac{142857}{999999} = \frac{1}{7}$$

since 7×142857 equals 999999 exactly.

You can do this trick with

```
     r = 0.333
10 x r = 3.33
 9 x r = 3
```

$$r = \frac{3}{9} = \frac{1}{3}$$

or even

```
     r = 0.4999
10 x r = 4.999
 9 x r = 4.5
```

$$r = \frac{4.5}{9} = \frac{1}{2}$$

and

```
     r = 0.9999
10 x r = 9.999
 9 x r = 9
```

$$r = \frac{9}{9} = 1$$

This is one of the subtleties of numbers. In what sense can we say that

$$0.5 = 0.4999\ldots$$

$$1 = 0.9999\ldots$$

Most everyone is OK with the example $1/3 = 0.3333\ldots$ but some may be uneasy with the other two.

Ultimately, we justify the result as defined by evaluation of a limit.

Consider 0.9999. If n is the number of places in the result, then as $n \to \infty$ the number being shown approaches 1 as its limit. We'll come back to this after considering the real numbers.

ordering

For two rational numbers a and b there are only three cases:

$$r_1 = r_2, \qquad r_1 < r_2, \qquad r_1 > r_2$$

It is a property of the integers, that if

$$a < b$$

then for $c > 0$

$$ca < cb$$

using that property

$$\frac{p}{q} < \frac{s}{t} \iff pt < qs$$

p/q is less than s/t if and only if $pt < qs$. (If only one of p and q or s and t is negative, associate the minus sign with the numerator).

Ordering of the integers guarantees ordering of the rational numbers. For any rational numbers, if

$$\frac{p}{q} < \frac{s}{t}$$

then for $c > 0$

$$c \cdot \frac{p}{q} < c \cdot \frac{s}{t}$$

For that matter, it is generally true for real numbers (which include integers and rationals) that if

$$a < b$$

for $c > 0$

$$ca < cb$$

intervals

We denote the numbers greater than u and less than v as lying in the interval (u, v). With parentheses, the interval described is *open*, it does not include the boundary values.

To describe a *closed* interval, write $[u, v]$. This interval includes all the values in the first one, plus it also includes u and v.

Because of the density property described below, any interval such as

$$I = [0, 1]$$

contains an *infinite* quantity of rational numbers.

density

Consider the set of all points

$$x = \frac{p}{10^n}$$

for all natural numbers n and integers p.

It is clear that simply by increasing the value of n, we can construct a set of equally spaced rational numbers as tightly clustered as we wish.

The rational numbers are said to be *dense* on the number line.

Chapter 31

Infinity

infinity

The symbol for infinity is ∞.

In the old days, they used to write things like

$$\frac{1}{0} = \infty, \qquad \frac{1}{\infty} = 0$$

John Wallis wrote $24/0 = \infty$, in 1656, which is when the ∞ symbol was introduced with its current definition. Even Euler argued that $n/0 = \infty$ when it suited him,

It is claimed that the symbol derives from the Roman symbol for 100 million. That's interesting. I never knew any symbols larger than M, for one thousand. And I'm not sure I believe it, but that's what some people say.

According to

https://notevenpast.org/dividing-nothing/

> On 21 September 1997, the USS Yorktown battleship was testing "Smart Ship" technologies on the coast of Cape Charles, Virginia. At one point, a crew member entered a set of data that mistakenly included a zero in one field, causing a Windows NT computer program to divide by zero. This generated an error that crashed the computer network, causing failure of the ship's propulsion system, paralyzing the cruiser for more than a day.

no division by zero

There is a fundamental problem when we set up a division problem and 0 is in the denominator. What goes wrong when we attempt to divide by zero?

$$\frac{a}{0} = ?$$

Well, what do we mean by an expression such as

$$\frac{a}{b} = c$$

By *definition*, we mean that we will try to find c such that

$$c \cdot b = a$$

For the integers, of course, there is the problem of a possible remainder. Let us leave that aside for a minute.

Suppose we have $c \cdot b = a$ but then take b to be very small though not 0. In that case, the number c may get very large. That's OK.

We can make b as small as we wish by making c large enough or vice versa. And we can say that as $b \to 0$, then $c \to \infty$.

But we can't say $a/0 =$ some number.

If there were such a number (say $a/0 = \infty$, infinity), then what about

$$\frac{b}{0} = ??, \qquad \frac{c}{0} = ??$$

It would mean that whatever the expression $b/0$ is equal to, when multiplied by zero, we would obtain any number whatsoever. This makes no sense.

Here is another, perhaps silly, example.

$$0 \cdot 1 = 0$$

$$0 \cdot 2 = 0$$

so

$$0 \cdot 1 = 0 \cdot 2$$

but then

$$1 = 2$$

By definition, we do not allow division by zero.

infinity is not a number

And we can't answer the question what is $2 \cdot \infty$? If we allowed multiplication by ∞ then the only reasonable answer would be

$$2 \cdot \infty = \infty$$

so then also

$$n \cdot \infty = \infty$$

where n is any number. But then say

$$2 \cdot \infty = 3 \cdot \infty$$

so, cancelling

$$2 = 3$$

This would be a mess.

By definition, *infinity is not a number* and division by 0 is *undefined.*

limits

Often people say that calculus is all about limits, and they are certainly where you start in proving the theoretical basis of the field.

We will keep the discussion of limits and ϵ-δ formalism to a minimum for the reasons explained in the Introduction. But let us try to establish an intuitive idea about what we mean when we say "in the limit as $N \to \infty$".

Above we had that there is no greatest integer.

A corollary of that is the limit

$$\lim_{n \to \infty} \frac{(n+1) - n}{n} = 0$$

Why? As n increases without bound, the difference between successive numbers, as a fraction of n, tends to zero.

To get an idea about this, first simplify by multiplying by $1/n$ on top and bottom. Then we have

$$\lim_{n \to \infty} \frac{(1 + 1/n - 1)}{1} = \frac{1}{n}$$

We say that $1/n$ *tends* to zero as $n \to \infty$, and so does $[(n+1) - n]/n$.

Chapter 32

Euclid's algorithm

Consider two natural numbers a and b. Usually a is allowed to be an integer (i.e., it can be negative), but to keep things simple here we will say that $a, b \in \mathbb{N}$, a and b are positive integers.

We can find their *greatest common divisor*, written (a, b). First we write the unique prime factorization of a and b:

```
180 =           2 x 2 x 3 x 3 x 5
140 =           2 x 2 x         5 x 7
gcd(140,180) = 2 x 2 x         5 = 20
```

Pick out the common factors and the $\gcd(a, b)$ will be their product. It is important that we do not need to actually factor a and b.

(We will develop a theorem on unique prime factorization in another chapter).

The algorithm works like this. Find integers $r \geq 0$ and $q > 0$ such that

$$a = b \cdot q + r$$

• If $r = 0$ we are done: b divides a equally. Otherwise

∘ switch $a = b$ and $b = r$ and repeat.

Then b is the gcd of the original a and b.

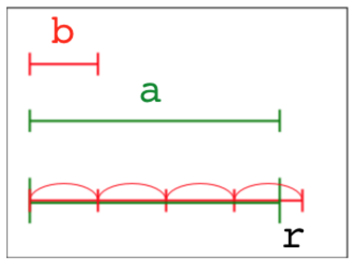
In our example

```
180 = 140 x 1 + 40
140 =  40 x 3 + 20
 40 =  20 x 2 + 0
gcd =  20
```

Here is the reason this works. First, we can always find q and r such that

$$a = b \cdot q + r$$

This is a version of the Archimedean property for positive integers.



It may be paraphrased by saying

> given a bathtub full of water and a teaspoon, it is possible to empty the bathtub.

Either $a = b \cdot q$ and we are done or:

$$b \cdot q < a < b \cdot q + b$$

So then

$$a - bq > 0$$

$$a - bq < b$$

With $r = a - bq$, we obtain $0 < r < b$.

Let u be the largest integer that divides both a and b (the greatest common divisor)

$$a = su$$

$$b = tu$$

Then

$$su = q \cdot tu + r$$

$$r = su - q \cdot tu$$

$$r = u(s - q \cdot t)$$

So u divides r.

Hence every common divisor of a and b is also a divisor of b and r.

recursive program

Here are two examples of programs in different styles that implement the algorithm (with no error checking):

```
def gcd(a,b):
    r = a % b
    if r == 0:
        return b
    return gcd(b,r)
```

```
def gcd(a,b):
    r = a % b
    while r != 0:
        a,b = b,r
        r = a % b
    return b
```

The first version is *recursive*, it may call itself. The second uses a **while** loop to accomplish the same thing.

Chapter 33

Real numbers

There is a big problem with rational numbers which you probably know: some numbers cannot be expressed as the ratio of two integers, as a first example, the number which when multiplied by itself is equal to 2, written $\sqrt{2}$.

The discovery that one cannot find integer p and q such that

$$(\frac{p}{q})^2 = 2$$

is due to the Pythagorean school and was most unwelcome since it screwed up their cherished theory of the universe.

Some say that they drowned the guy who discovered it by throwing him overboard, and that his name was Hippasus. Like most stories about Greek mathematicians, the truth is unknown.

We will see that there is a similar problem (called irrationality) with $\sqrt{3}$, $\sqrt{5}$, $\sqrt{7}$, etc., as well as with $3^{1/3}$ and so on.

Proof.

For $\sqrt{2}$:

We assume that there does exist a rational number p/q such that

$$\frac{p}{q} = \sqrt{2}$$

We will show that this assumption leads to a contradiction.

A crucial part of the proof is that we suppose p/q to be in lowest terms and in particular, that p and q are not both even. It would be easy to recognize the case if they were both even, for then each would have their terminal digit in the set $\{0, 2, 4, 6, 8\}$.

Another fact we will need is that every odd number, when squared, gives an odd result. Proof: every odd number can be written as $2k + 1$ (for non-negative integer k) and then

$$(2k + 1)^2 = 4k^2 + 4k + 1$$

which is an odd number. Therefore, if n^2 is even, n is also even.

So go back to

$$\frac{p}{q} = \sqrt{2}$$

Move the q term to the right-hand side and square both sides:

$$p^2 = 2q^2$$

This implies that p^2 and p are even, using the result from above. So we can write that $p = 2m$. But now

$$(2m)^2 = 2q^2$$
$$2m^2 = q^2$$

which implies that q is *also* even.

We started with the assumption that p and q are not both even, but now we've reached a contradiction. We conclude that there do not exist two integers p and q such that $p/q = \sqrt{2}$.

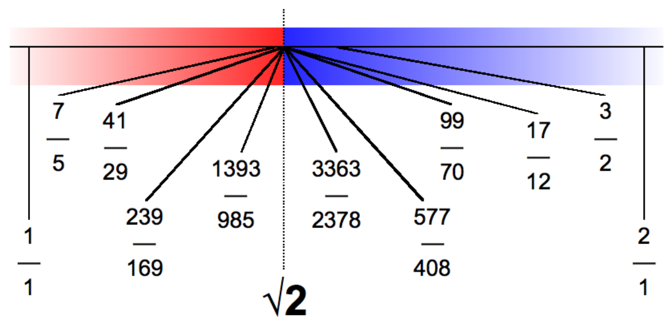
discussion

To quote Hardy (*A Mathematician's Apology*):

> The proof is by reductio ad absurdum, and reductio ad absurdum, which Euclid loved so much, is one of a mathematician's finest weapons. It is a far finer gambit than any chess gambit: a chess player may offer the sacrifice of a pawn or even a piece, but a mathematician offers *the game*.

The numbers like $\sqrt{2}$ are said to be *irrational* numbers and the set of these, plus all the other numbers is called the set of real numbers $\mathbb{R}$.

This led Dedekind to formulate the famous Dedekind cut. Visualize the standard number line as an infinite line on (an infinite) piece of paper.



Each real number corresponds to a cut, a knife-edge coming down somewhere on this number line. Every other number that is not equal to this one, is either $>$ or $<$ the number specified by the cut.

One position is $\sqrt{2}$, another is $3/2$ and so on.

proof using prime factors

The fundamental theorem of arithmetic says that any positive integer greater than 1 can be expressed as a product of its prime factors

$$n = p_1 \cdot p_2 \ldots p_k$$

where this factorization is unique (if the factors are sorted first), and multiple copies allowed. For example

$$60 = 2 \cdot 2 \cdot 3 \cdot 5$$

A corollary says that the square of any integer (a perfect square) has an even number of prime factors since

$$n^2 = p_1^2 \cdot p_2^2 \ldots p_k^2$$

In the expression from above

$$p^2 = 2q^2$$

the number of prime factors on the left is therefore even, but the number on the right is odd. This is a contradiction. Therefore p and q cannot both be integers.

continued fractions

Square roots can be represented as continued fractions. Some smart person figured out that we can write this:

$$(\sqrt{2}-1)(\sqrt{2}+1) = 2-1 = 1$$

Now, rearrange to get a substitution we will use repeatedly

$$\sqrt{2}-1 = \frac{1}{\sqrt{2}+1}$$

Add one and subtract one on the bottom right:

$$\sqrt{2}-1 = \frac{1}{2+\sqrt{2}-1}$$

And substitute for $\sqrt{2}-1$:

$$= \frac{1}{2+\frac{1}{\sqrt{2}+1}}$$

Lather, rinse, and repeat:

$$= \frac{1}{2+\frac{1}{2+\sqrt{2}-1}} = \frac{1}{2+\frac{1}{2+\frac{1}{\sqrt{2}+1}}}$$

Clearly, this goes on forever.

$$\sqrt{2}-1 = \cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{2+\dots}}}$$

Add 1 to the value of the *continued fraction* to get an expression for the square root of 2.

The numerators are all 1, so this is called a simple continued fraction. The continued fraction representation of $\sqrt{2}$ is usually written as [1 : 2], meaning that there is an initial 1 followed by repeated 2's.

This fraction goes on forever (since $\sqrt{2}$ is irrational). One can view the existence of the infinite continued fraction as a proof of irrationality.

We can turn the above into an approximate decimal representation of $\sqrt{2}$, by truncating the infinite expansion at the .... Then the last fraction is 5/2. Invert and add, repeatedly:

```
2 + 1/2 = 5/2
2 + 2/5 = 12/5
2 + 5/12 = 29/12
2 + 12/29 = 70/29
2 + 29/70 = 169/70
2 + 70/169 = 408/169
```

To terminate we need to use that initial 1:

```
1 + 169/408 = 577/408 = 1.414216
```

To six places, $\sqrt{2} = 1.414213$. We have five places, and can easily get more.

geometric proof

There are many other proofs of the irrationality of the square root of 2.

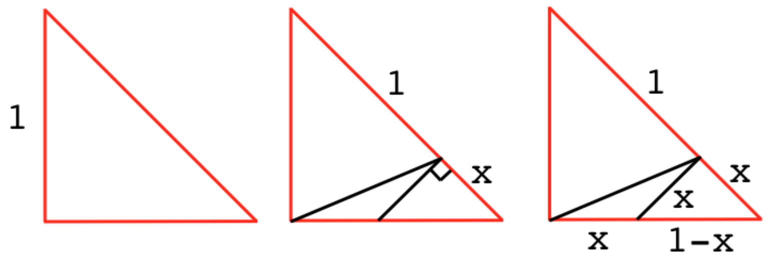
https://www.cut-the-knot.org/proofs/sq_root.shtml

Here we will look at one more, before considering a more general proof for all non-perfect squares. This one is from Tom Apostol (see the link). A more elaborate exposition is:

https://jeremykun.com/2011/08/14/the-square-root-of-2-is-irrational-geometric-proof/

Draw an isosceles triangle with side length 1, then Pythagoras tells us that the hypotenuse is equal in length to $\sqrt{2}$ (left panel).

Our hypothesis is that this length is a rational number, and its ratio to the side is in "lowest terms".



Mark off the length of the side (length 1) on the hypotenuse, and erect a perpendicular (middle panel). Also draw the line segment to the opposite vertex of the original triangle.

The new small triangle that is formed containing the right angle and with side length x in the middle panel is isosceles, because it is a right triangle, and it contains one of the complementary angles of the original right triangle.

By hypothesis, its side length x is the difference of two rational numbers, so x is a rational number.

Furthermore, the *other* small triangle is also isosceles. Its base angles, when added to the equal angles of an isosceles triangle, form right angles. This allows us to mark the side along the base as having length x as well.

Therefore, the hypotenuse of the new, small right triangle is a rational number, since it is equal to $1 - x$.

We are back where we started, with an isosceles triangle that has all rational sides.

It is clear that this process can continue forever. The sides will never be in "lowest terms" because we can always form a new similar but smaller right triangle, which amounts to evenly dividing both the sides and the hypotenuse by a rational number.

general proof

I found a long algebraic proof of the general irrationality of roots and I wrote it up the big calculus book. But then I came upon simple elegant proof based on the fundamental theorem of arithmetic.

We suppose that there exist two integers a and b such that

$$(\frac{a}{b})^2 = n$$

Both a and b have a unique prime factorization. Suppose that gives $a = a_1 \cdot a_2 \dots a_i$ and likewise for b so:

$$(\frac{a_1 \cdot a_2 \dots a_i}{b_1 \cdot b_2 \dots b_j})^2 = n$$

There must be at least some b_j which are not a_i, otherwise we could cancel all of them and so a/b would be an integer.

Call that factor or factors q, so in lowest terms we have

$$(\frac{a_1 \cdot a_2 \dots}{q_1 \dots q_k})^2 = n$$

But then, after squaring, we will have q_1^2 (for example) in the denominator and no corresponding factor of either q_1 or q_1^2 in the numerator.

The q_k cannot be canceled and so the result cannot be an integer.

This proves that the only n with rational square roots are perfect squares with integer roots.

The proof also applies generally to other powers like cube and the fourth and fifth power and so on.

other irrational numbers

There are many other irrational numbers besides these square roots.

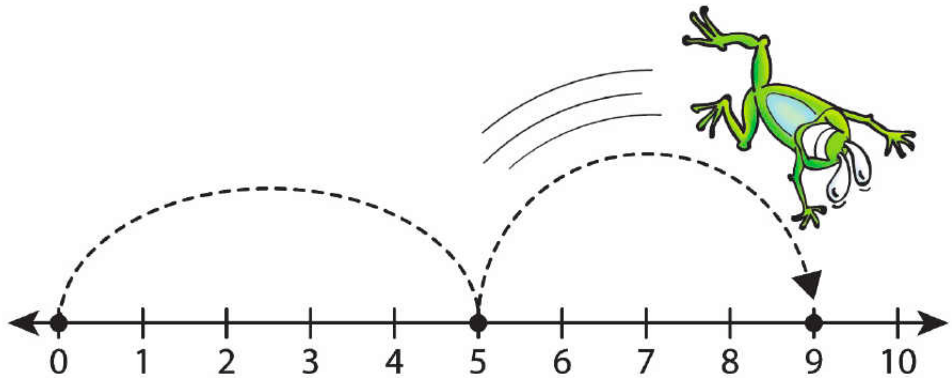
The proof that e is irrational is easy, but since we haven't introduced the exponential yet we need to wait. The proof that π is irrational is bit harder, we will skip that.

density

number line

A simple tool to visualize all of the real numbers is the familiar number line. Here is the number line with numbers marked from $\mathbb{N}$, but obviously we could also draw one for $\mathbb{Z}$ or $\mathbb{Q}$.

We explore the application of the number line to $\mathbb{R}$ as we proceed.



We might simply assume that to every point on the number line there corresponds a rational or irrational number, and that this total collection obeys the same laws of arithmetic as the rational numbers do.

As mentioned above, the need for the real numbers is indicated by empty "holes" in the number line corresponding to the irrational numbers like $\sqrt{2}$.

A problem that arises is how to specify an irrational number non-geometrically and other than as the solution to an equation such as $r^2 = 2$. We saw above a method involving continued fractions.

approximations

In all cases we write particular real numbers as *approximations*. For example, the square root of 2 lies between 1 and 2 because

$$1^2 = 1 < 2$$

$$2^2 = 4 > 2$$

Implying that $\sqrt{2} < 2$. At the second place:

$$1.4^2 = 1.96 < 2$$

$$1.5^2 = 2.25 > 2$$

Implying that $\sqrt{2} < 1.5$. At the third:

$$1.41^2 = 1.9881 < 2$$

$$1.42^2 = 2.0164 > 2$$

Implying that $\sqrt{2} < 1.42$.

This process may be continued for as long as desired.

We can never write down the decimal value of $\sqrt{2}$ exactly, but only approximate it to greater and greater precision. It goes on forever.

In carrying out this recursive process, suppose we know 1.41 and we seek the next digit. Rather than try all the digits in order starting with 1, there is a better way.

Try to estimate the error from the previous round.

For example $1.41^2 = 1.9881$ so we are short of 2.0000 by 0.0119.

$1.42^2 = 2.0164$ so the difference is 0.0283 and the fraction of the difference that we're under is $119/283 = 0.4205$. In fact, the next two digits of the approximation to $\sqrt{2}$ are 42.

However, we will see a much better method for obtaining this value later, called Newton's method.

At the seventh place

$$1.414213^2 = 1.9999984093689998.. < 2$$

$$1.414214^2 = 2.0000012377960004 > 2$$

Because any repeating decimal can be written as a fraction, we know that the sequence cannot repeat (any apparent repeat will be illusory).

It is a curious fact that all the digits of π, *to whatever accuracy you desire*, can be found in the correct order, somewhere within the digital expansion of e or ϕ or indeed, any irrational number. The converse is also true.

Another way to say the same thing is that *any* finite sequence can be found within *any* infinite sequence, and in as many copies as you have the patience to discover. The sequence 271828 is found starting around digit 33,790 of π, but 2718281 (adding the next digit of e) is not found within the first million digits of π. You just need more.

limit of a sequence

The real number $\sqrt{2}$ is defined to be the limit of the sequence

$1.4, 1.41, 1.414, \ldots 1.414214\ldots$

as the number of terms $n \to \infty$.

In a similar way, the number e can be viewed as

$$\lim_{n\to\infty}(1+\frac{1}{n})^n$$

And the number π can be viewed as the limit of the method of exhaustion applied to the area of a unit circle.

density of numbers

We showed previously that between any two rational numbers, including 0 and the *smallest* positive number, one can find another rational number which lies between them.

Three related statements are also true.

- for any two rational numbers one can find a real number which lies between them
- for any two real numbers one can find a rational number which lies between them
- for any two real numbers one can find a real number which lies between them

Proofs of these are readily accessible but we'll do only one, for the sake of brevity.

theorem

- Between any two rational numbers it is always possible to find a real number.

Proof.

We will find a *particular* irrational in the interval (a, b), where a and b are rational. For $a < b$, we simply add to the number a the following

$$c = \frac{b-a}{\sqrt{2}} = \frac{\sqrt{2}}{2}(b-a)$$

c is smaller than $b-a$ (because $\sqrt{2}/2 < 1$) so the result $a+c$ lies between a and b.

We also know that c is irrational, because $\sqrt{2}$ times any rational number is irrational. Finally, $a+c$ is irrational because adding $\sqrt{2}$ times a rational number to any rational number produces an irrational number.

Proof of the first preliminary requirement: $\sqrt{2}$ times a rational is irrational. Suppose for integer p, q, r, s we have

$$\sqrt{2}\,\frac{p}{q} = \frac{r}{s}$$

then

$$\sqrt{2} = \frac{rq}{ps}$$

But the right-hand side is rational, so this is a contradiction.

For the second requirement, again by contradiction suppose

$$\sqrt{2}\,\frac{p}{q} + \frac{s}{t} = \frac{u}{v}$$

for integer p, q, r, s, u, v. But the right-hand side of

$$\sqrt{2} = \frac{q}{p}(\frac{u}{v} - \frac{s}{t})$$

is rational, so this is a contradiction.

□

Note in passing that powers are different. What do you think about

$$r = \sqrt{2}^{\sqrt{2}}$$

You may think r is "likely" to be irrational. Just a mess. But how about

$$r^{\sqrt{2}}$$

Whether r is rational or irrational

$$r^{\sqrt{2}} = (\sqrt{2}^{\sqrt{2}})^{\sqrt{2}} = \sqrt{2}^2 = 2$$

This property of the real (and even the rational) numbers, that there is no closest number to any given number, accounts for virtually all of the theoretical difficulties in calculus which are solved by the use of limits and the apparatus of δ and ϵ or alternatively, neighborhoods. We will get to that in a bit.

Part VIII

Algebra

Chapter 34

Basic algebra

There is very little algebra that needs to be memorized for where we're going. We just finished a chapter about the sum of integer squares. If you can follow that, you're in good shape. If not, go back and work through it again, carefully.

Here is a bit more:

inequality

You have surely seen and used the symbols $>$ (greater than), and $<$ (less than) before we used them a second ago.

Among the axioms of the number systems is the collection of *order axioms*. A few definitions:

- $x < y$ means that $y - x$ is positive
- $y > x$ means that $x < y$

For arbitrary numbers a and b only one of three statements is true:

- $a < b$
- $a = b$
- $a > b$

There is no attempt or need to be systematic here. Let us just mention that these properties (and their kin) are true not just for natural numbers, but also for the rational numbers and the real numbers, as we will see in due course.

Here are a few important theorems about order which we will use often:

- If $a < b$, and c is any number, then $a + c < b + c$
- If $a < b$, then $-b < -a$
- If $a < b$ and $c > 0$, then $ac < bc$

The first one above implies the second and the third.

algebraic operations

- addition: $a + b$
- subtraction: $a - b = a + (-b)$

The negative integers and 0 solve the problem of how to evaluate $a - b$ when $b \geq a$.

- multiplication: $a \cdot b$, also often written ab (but not $a \times b$, at this level).

And then:

- division a/b, equivalent to finding a number c such that $c \cdot b = a$.

algebra

As you know, the basic axioms of algebra include the following:

- Commutativity for addition and multiplication:

$$a + b = b + a, \qquad a \cdot b = b \cdot a$$

- Associativity for addition and multiplication:

$$(a + b) + c = a + (b + c), \qquad (a \cdot b) \cdot c = a \cdot (b \cdot c)$$

- Distributivity of addition over multiplication:

$$a \cdot (b + c) = a \cdot b + a \cdot c$$

- Additive identity: $0 + a = a$.
- Multiplicative identity: $1 \cdot a = a$.

binomial theorem

One basic idea from algebra is the binomial theorem:

$$(a+b)^2 = (a+b)(a+b) = a^2 + 2ab + b^2$$

so then

$$\begin{aligned}(a+b)^3 &= (a+b)(a^2 + 2ab + b^2)\\ &= a^3 + 3a^2b + 3ab^2 + b^3\end{aligned}$$

And in general, to get the next $n+1$ power, go through the expansion for the n power and multiply each term separately by a and b.

You will find that the cofactors are given by Pascal's triangle.

$$\begin{array}{ccccccccccc} &&&&&1&&&&& \\ &&&&1&&1&&&& \\ &&&1&&2&&1&&& \\ &&1&&3&&3&&1&& \\ &1&&4&&6&&4&&1& \\ 1&&5&&10&&10&&5&&1 \end{array}$$

$$\begin{aligned}(a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3\\ (a+b)^4 &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4\\ (a+b)^5 &= a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5\\ &\dots\end{aligned}$$

If you substitute $-b$ for b you will find that everything is exactly the same, except those terms with b raised to an odd power have acquired a minus sign.

$$(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$$

Hence

$$(a-b)(a^2 - 2ab + b^2) = a^3 - 3a^2b + 3ab^2 - b^3$$

The binomial theorem is usually stated and worked with in terms of positive integers.

But it actually works for negative integers and fractional powers as well. The big difference is that the series of terms is *infinite*.

Newton discovered that, though he didn't prove it. He just used it as a tool. He found that

$$\frac{1}{1+x} = (1+x)^{-1}$$
$$= 1 - x + x^2 - x^3 + x^4 + \cdots$$

Newton checked this by multiplying:

$$(1+x)(1 - x + x^2 - x^3 + x^4 + \cdots)$$
$$= 1 - x + x^2 - x^3 + x^4 + \cdots$$
$$+x - x^2 + x^3 - x^4 + \cdots = 1$$

But be careful! What happens if $x = -1$?

factoring

Here's something we will use often: the difference of two squares.

$$a^2 - b^2 = (a+b)(a-b)$$

The classic quadratic equation is often written

$$y = ax^2 + bx + c$$

This is a parabola that opens up ($a > 0$).

Depending on the values of the *coefficients* a, b, c, this equation may or may not have solutions when $y = 0$

$$ax^2 + bx + c = 0$$
$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0$$

Suppose that r and s are such solutions, then

$$(x-r)(x-s) = 0$$

and so

$$x^2 - (r + s)x + rs = 0$$

A fair amount of effort in algebra goes into guessing values of r and s that work, that have the appropriate sum and difference to match the equation we are given.

However, the answers frequently are not integers, and in that case, this approach is doomed.

A formula that always works is the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

I said "always works". To be more precise, it works if there are any solutions. If the discriminant $D = b^2 - 4ac$ is negative, then we're trying to take a square root of a negative number, and we don't know how to do that in the real numbers.

We will develop this more in the chapters on analytic geometry. That's really all you will need.

Chapter 35

Exponential

Principal and interest

Suppose I put 100 dollars in the bank, and the people at the bank say that after one year, they will give me an additional \$10 at that time. They will pay 10% interest for the year on the principal P of \$100.

However, suppose I bargain with them. I get them to promise to pay me half the interest (5%) at the six-month mark, and the rest after one year. My account will hold \$105 after six months, and the interest due for the second half will be 5% of \$105, which is \$5.25 for a total of \$10.25.

The equation to describe this situation is that if the rate of interest for the year is r and the year is broken up into n periods when interest will be paid, the total amount at the end will be:

$$A = P(1 + \frac{r}{n})^n$$

In the example, we have $r = 0.10$ and $n = 2$ so

$$A = 100(1 + 0.05)^2 = 110.25$$

This is compound interest. If there are additional years t, the exponent will be nt rather than n.

And now we start wondering what happens if the bank pays every month so that $n = 12$ or every day so $n = 365$ or even every second. What happens if the interest is compounded *continuously*?

$$A = \lim_{n\to\infty} P\ [\ (1 + \frac{r}{n})^n\ ]$$

Now it turns out that in the limit as n approaches ∞ these two expressions are equal

$$(1 + \frac{r}{n})^n = \ [\ (1 + \frac{1}{n})^n\ ]^r$$

The same factor r can be either in the numerator of the second term inside or up in the exponent outside.

A quick proof is:

$$\lim_{n\to\infty} (1 + \frac{r}{n})^n$$

$$= \lim_{n\to\infty} (1 + \frac{r}{n})^{(n/r)r}$$

Define $m = n/r$ and so as $n \to \infty$, so does $m \to \infty$ and then we have

$$\lim_{m\to\infty} (1 + \frac{1}{m})^{(m)r}$$

and the r is outside. m is just a dummy variable so we write:

$$\lim_{n\to\infty}\ [\ (1 + \frac{1}{n})^n\ ]^r$$

□

Therefore, going back to what we were working on, let us bring out the factor r and obtain

$$A = P(1 + \frac{1}{n})^{nr}$$

$$A = P\ [\ (1 + \frac{1}{n})^n\ ]^r$$

Thus, the important question is, what is the value of this expression?

$$A = \lim_{n\to\infty} (1 + \frac{1}{n})^n$$

It does not depend on r. It will turn out that this limit is equal to the number e.

$$e = 2.71828\ 18284\ 59045\dots$$

That's really all we need to worry about with respect to e, for now.

As far as general exponents go, I'm sure you know that:

$$x^a \ x^b = x^{a+b}$$

$$(x^a)^e = x^{ae}$$

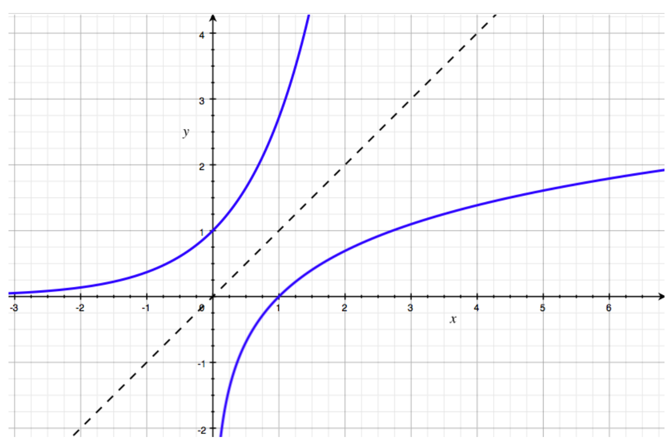
$$x^{-a} = \frac{1}{x^a}$$

$$x^{1/2} = \sqrt{x}$$

Chapter 36

Logarithms

The logarithm and the exponential are inverse functions, we can see that if we plot them together:



The upper curve is $y = e^x$ and the lower one is $y = \ln x$. As inverse functions, they are symmetric about the line $y = x$.

If we have that

$$y = b^x$$

for some $b > 0, b \neq 1$, then we say that

$$x = \log_b y$$

Putting them together

$$y = b^{\log_b y}$$

The usual bases are

- 10 (common logarithm, $\log_{10}$, or just log)
- e (natural logarithm or ln)
- 2 (binary logarithm, $\log_2$).

The rules for exponents are simple, if p and q are two numbers and we know the logarithms of p and q to base b

$$p = b^u$$

$$q = b^v$$

then their product can be computed as:

$$pq = b^u \cdot b^v = b^{u+v}$$

To multiply two numbers, *add* their logarithms.

It helps if we can actually compute b^{u+v}. In the old days there were tables of logarithms, so you just looked up the answer in the table.

LOGARITHMS, BASE 10 $\log_{10} x$ or $\lg x$

x	0	1	2	3	4	5	6	7	8	9
10	.0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	.0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	.0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	.1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	.1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	.1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	.2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	.2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	.2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	.2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	.3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	.3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	.3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	.3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	.3802	3820	3838	3856	3874	3892	3909	3927	3945	3962

$\log 2 \approx 0.3010$.

https://en.wikipedia.org/wiki/Mathematical_table#Tables_of_logarithms

The second rule is for exponentiation:

$$(b^u)^v = b^{uv}$$

And in terms of logarithms we can write

$$uv = \log_b (b^u)^v$$
$$= \log b^{uv} = v \log_b (b^u)$$

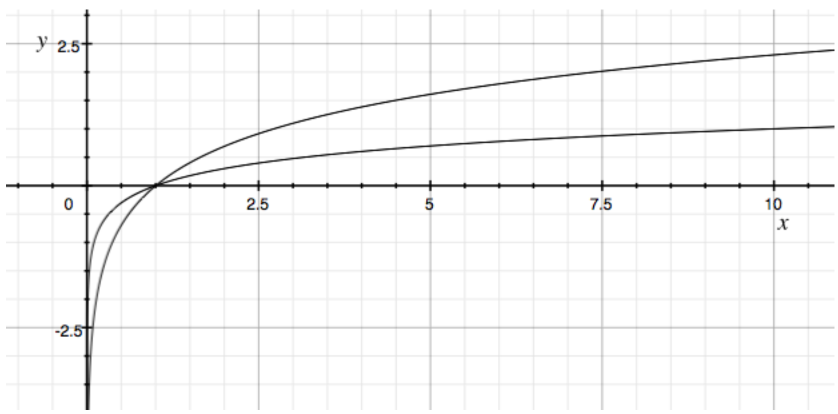
For example

$$2^2 = 2 \cdot 2 = 4$$
$$2^3 = 2 \cdot 2 \cdot 2 = 8$$
$$4 \cdot 8 = 2^2 \cdot 2^3 = 2^{2+3} = 2^5$$
$$= 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 = 32$$

and

$$(2^2)^3 = 4^3 = 64 = 2^6 = 2^{2 \cdot 3}$$

Here is a plot of $\log_{10}(x)$ and $\ln x$:



The first function reaches the value 1 when $x = 10$ and the second reaches the value 1 when $x = e$. Both have the value 0 at $x = 1$ because $b^0 = 1$ for any base, so the logarithm to any base of 1 is equal to 0.

It turns out that if we take the logarithm of x (where x is any number > 1) to two *different* bases, the ratio of the logarithms is a constant, independent of the value of x. And it is not hard to imagine that the ratios of the two values for any x is a constant, in the plot above.

change of bases

This relationship is nicely shown by the change of bases formula.

$$\log_b x = \frac{\log_a x}{\log_a b}$$

Derivation.

Start with an expression with b as the base:

$$y = b^x$$

and by the definition of the logarithm

$$x = \log_b y$$

To derive the formula, take the logarithm to the base a on both sides of the first expression:

$$\log_a y = \log_a(b^x)$$

Now, just invoke the second rule on the right-hand side

$$= x \log_a b$$

and then substitute for x from the second expression above

$$\log_a y = \log_b y \log_a b$$

We're basically done.

□

y can be any value, so replace it by x

$$\log_a x = \log_b x \log_a b$$

Rearranging:

$$\log_b x = \frac{\log_a x}{\log_a b}$$

Three ideas for remembering the formula.

(1) Learn the derivation.

(2) Logarithms of x to different bases b and a are connected by some constant k

$$\frac{\log_b x}{\log_a x} = k$$

$$\log_b x = k \log_a x$$

and we substitute for k the inverse of the log to the *same* base as we have in the numerator:

$$\log_b x = \frac{1}{\log_a b} \cdot \log_a x$$

that is, I remember that we want $\log_a$ something *over* $\log_a$ something on the right.

(3) You might look at the other formula

$$\log_a x = \log_a b \log_b x$$

and imagine the b's canceling in some way.

One other thing we can do is to set $x = a$ in the above formula. We start from

$$\log_b x = \frac{\log_a x}{\log_a b}$$

then with $x = a$

$$\log_b a = \frac{\log_a a}{\log_a b}$$

but $\log_a a = 1$ so

$$\log_b a = \frac{1}{\log_a b}$$

And that makes perfect sense. If we multiply by some factor k to convert from the logarithm in base a to base b, we must multiply by the inverse of the same factor to convert back again.

For the figure above of the common log (base 10) and the natural logarithm, $\ln 10 = 2.303$, and that looks about right, when $x = 10$ the first function is 1.0 and the second one is about 2.3.

fractional exponents

The introduction above dealt mainly with integer exponents, but of course you know that the practical use of logarithms depends on fractional values. The simplest way to see how this works is to consider the square root.

$$\sqrt{2} \times \sqrt{2} = 2$$

If we think about what the exponent u to the base 2 would be such that

$$2^u = \sqrt{2}$$

We observe that by the rules for exponents

$$\sqrt{2} \times \sqrt{2} = 2^u \times 2^u = 2^{u+u} = 2^1$$

That is

$$u + u = 1$$

so $u = 1/2$. By the same logic the n^{th} root of b is $b^{1/n}$. And of course

$$(b^2)^{1/2} = b^{2 \times 1/2} = b^1$$

fractional exponents

Feynman has a nice description of how logarithms were calculated (see Lectures, volume 1, Chapter 22, Algebra)

http://www.feynmanlectures.caltech.edu/I_22.html

The basic idea is to take repeated square roots of the base (10), and then combine those to form the required value.

So for example

$$10^{1/2} = 3.1622776602$$

this has been rounded up, the next term in the expansion is $3.162277\ldots$. To obtain $10^{1/4}$, compute the square root of $10^{1/2}$.

$$10^{1/4} = 1.7782794100$$

$$10^{1/8} = 1.3335214321$$

$$10^{1/16} = 1.1547819847$$
$$10^{1/32} = 1.07460782832$$
$$10^{1/64} = 1.03663292844$$
$$10^{1/128} = 1.0181517217$$
$$10^{1/256} = 1.009035044841$$
$$10^{1/512} = 1.004507364254$$
$$10^{1/1024} = 1.002251148293$$

and so on. Eventually (around 10^{1024}) we stop, because there is a neat trick that saves us from calculating. It's a result from basic calculus so we won't say more.

Having these powers of 10, now we want to compute a logarithm, say $\log_{10} 2$. The first thing is that 2 is smaller than $10^{1/2}$.

So we start with

$$2 = 10^{1/4} \cdot ??$$

By trying the various powers of 10, we settle on

$$2 = 10^{1/4} \cdot 10^{1/32} \cdot 10^{1/64} \cdot 10^{1/256} \cdot r$$

We won't worry about the r, it is very close to 1.

To get the logarithm of 2 just add the logs of those powers:

$$\log_{10} 2 = 0.25 + 0.03125 + 0.015625 + 0.00390625 = 0.30078125$$

The actual value is 0.30103, to five places.

You can see that to get good accuracy, we will need to figure out the correction term.

Less than 1

Fractional exponents leads to consideration of $0 < x < 1$. Write

$$x \cdot \frac{1}{x} = 1$$

Take the logarithm of both sides

$$\log(x \cdot \frac{1}{x}) = \log 1 = 0$$

$$= \log x + \log \frac{1}{x}$$

Thus

$$\log \frac{1}{x} = -\log x$$

Chapter 37

Sum of squares

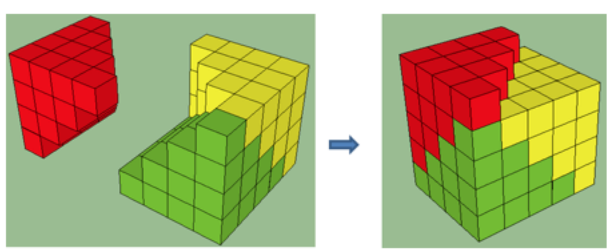
We want to find a formula for the sum of the squares of the first n integers, which is written variously as

$$\frac{n\ (n+1)(2n+1)}{6}, \qquad \frac{n(n+1)}{2}\cdot\frac{(2n+1)}{3}$$
$$\frac{1}{6}\ (2n^3+3n^2+2n), \qquad \frac{n^3}{3}+\frac{n^2}{2}+\frac{n}{3}$$

and my favorite:

$$\frac{1}{3}\ n\cdot(n+1)(n+\frac{1}{2})$$

which explains the proof without words:



sum of squares

We us use exactly the same method as for the sum of integers to determine the formula for

$$S_n = 1^2 + 2^2 + \dots n^2$$

Since the formula for the sum of integers has a square, we expect that this will be a cubic. Write

$$(k+1)^3 = k^3 + 3k^2 + 3k + 1$$

$$\sum_{k=1}^{n}(k+1)^3 = \sum_{k=1}^{n} k^3 + \sum_{k=1}^{n} 3k^2 + \sum_{k=1}^{n} 3k + \sum_{k=1}^{n} 1$$

Moving the first term from the right-hand side to the left, we obtain a telescoping sum as the difference, just like before:

$$(n+1)^3 - 1 = \sum_{k=1}^{n} 3k^2 + \sum_{k=1}^{n} 3k + \sum_{k=1}^{n} 1$$

Now it's just some messy algebra. The second term on the right-hand side includes our previous result:

$$n^3 + 3n^2 + 3n = 3\sum_{k=1}^{n} k^2 + 3\ \frac{n(n+1)}{2} + n$$

$$6\sum_{k=1}^{n} k^2 = 2(n^3 + 3n^2 + 3n) - 3n(n+1) - 2n$$

$$6\sum_{k=1}^{n} k^2 = n\ [\ 2(n^2 + 3n + 3) - 3(n+1) - 2\ ]$$

$$6\sum_{k=1}^{n} k^2 = n\ [\ 2n^2 + 3n + 1\ ]$$

$$6\sum_{k=1}^{n} k^2 = n\ (n+1)(2n+1)$$

$$\sum_{k=1}^{n} k^2 = \frac{n\ (n+1)(2n+1)}{6}$$

which can be re-written in various ways including:

This formula is also written variously as

$$\frac{n(n+1)}{2} \cdot \frac{(2n+1)}{3}$$

$$\frac{1}{6}\ (2n^3 + 3n^2 + 2n)$$

$$\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{3}$$

But I find it easiest to remember the first version.

We can check it by induction. The base case is easy

$$\frac{1(2)(3)}{6} = 1$$

Now for the induction step:

$$\frac{n(n+1)(2n+1)}{6} + (n+1)^2$$

$$= \frac{n+1}{6}\ [\ (n)(2n+1) + 6(n+1)\ ]$$

Look at what's in the brackets

$$(n)(2n+1) + 6(n+1)$$

$$= 2n^2 + 7n + 6$$

$$= (n+2)(2n+3)$$

$$= (n+1+1)(2(n+1)+1)$$

So altogether we have

$$= \frac{(n+1)(n+1+1)(2(n+1)+1)}{6}$$

which indeed, is the formula we had above, substituting $n+1$ for n.

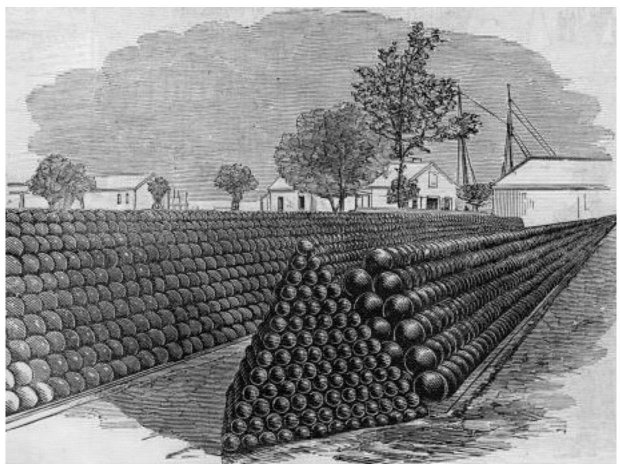
$\square$

Strang's proof

Here is another approach, from Strang's *Calculus*. He says "the best place to start is a good guess". So again, our goal is to find a formula for:

$$S = \sum_{k=1}^{n} k^2$$

Perhaps we visualize a pile of cannonballs



Each layer contains a square number of cannonballs (1, then 4, then 9, etc.). The shape is a pyramid with dimensions $n \times n \times n$. We know the formula for the volume of a pyramid, and guess

$$S_n = \frac{1}{3}n^3$$

To test it, check whether this difference is n^2 (as it should be):

$$S_n - S_{n-1} = \frac{1}{3}n^3 - \frac{1}{3}(n-1)^3$$

Now

$$(n-1)^2 = n^2 - 2n + 1$$
$$(n-1)^3 = (n-1)(n^2 - 2n + 1)$$

$$= n^3 - 3n^2 + 3n - 1$$

So

$$S_n - S_{n-1} = \frac{1}{3}(n^3 - n^3 + 3n^2 - 3n + 1)$$

We see that our guess is off by the residual terms

$$\frac{1}{3}(3n^2 - 3n + 1)$$
$$= n^2 - n + \frac{1}{3}$$

Strang says: the guess needs *correction terms.* To cancel $1/3$ in the difference, subtract $n/3$ from the sum. And to add back n in the difference, add back $1 + 2 + \cdots + n(n+1)/2$ to the sum. Our new guess is

$$S_n = \frac{1}{3}n^3 + \frac{n(n+1)}{2} - \frac{n}{3}$$
$$= \frac{n}{6}(2n^2 + 3(n+1) - 2)$$
$$= \frac{n}{6}(2n+1)(n+1)$$
$$= \frac{n(n+1)(2n+1)}{6}$$

which may be easier to remember as

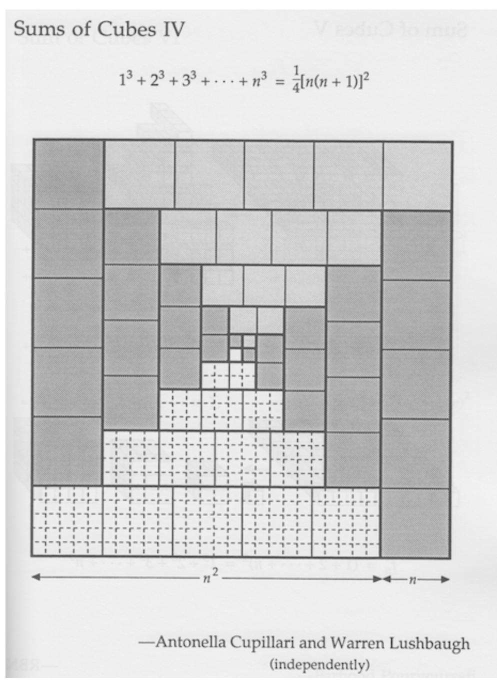
$$S_n = \frac{n(n+1)}{2} \cdot \frac{2n+1}{3}$$

Chapter 38

Sum of cubes

The formula is

$$\sum_{k=1}^{n} k^3 = [\sum_{k=1}^{n} k\ ]^2$$



Let's prove this using induction.

The "base case" is pretty simple. For $n = 2$

$$1^3 + 2^3 = 1 + 8 = 9$$

and

$$\frac{n^2(n+1)^2}{2^2} = \frac{2^2(3^2)}{2^2} = 3^2 = 9$$

Now for the induction step what we need to show is that what we get assuming the formula for n is correct and then adding the term $(n+1)^3$

$$\frac{n^2(n+1)^2}{2^2} + (n+1)^3$$

is equal to what we get by plugging $n+1$ into the formula.

$$\frac{(n+1)^2(n+2)^2}{2^2}$$

We need to show that eqn 2 is equal to eqn 3.

$$\frac{n^2(n+1)^2}{2^2} + (n+1)^3 = \frac{(n+1)^2(n+2)^2}{2^2}$$

First, we can factor out and cancel $(n+1)^2$ from both sides. So then we have

$$\frac{n^2}{2^2} + (n+1) \stackrel{?}{=} \frac{(n+2)^2}{2^2}$$

$$n^2 + 4(n+1) \stackrel{?}{=} (n+2)^2$$

That looks correct!

$\square$

derivation by collapsing sum

We proceed exactly as before

$$(k+1)^4 = k^4 + 4k^3 + 6k^2 + 4k + 1$$

Sum each term from $k = 1 \rightarrow k = n$

$$\sum_{k=1}^{n}(k+1)^4 = \sum_{k=1}^{n} k^4 + \sum_{k=1}^{n} 4k^3 + \sum_{k=1}^{n} 6k^2 + \sum_{k=1}^{n} 4k + \sum_{k=1}^{n} 1$$

Rearrange and compute the collapsing sum.

$$\sum_{k=1}^{n}(k+1)^4 - \sum_{k=1}^{n} k^4 = \sum_{k=1}^{n} 4k^3 + \sum_{k=1}^{n} 6k^2 + \sum_{k=1}^{n} 4k + \sum_{k=1}^{n} 1$$

$$(n+1)^4 - 1 = \sum_{k=1}^{n} 4k^3 + \sum_{k=1}^{n} 6k^2 + \sum_{k=1}^{n} 4k + \sum_{k=1}^{n} 1$$

Substitute for the right-hand sum

$$(n+1)^4 - 1 = \sum_{k=1}^{n} 4k^3 + \sum_{k=1}^{n} 6k^2 + \sum_{k=1}^{n} 4k + n$$

Rearrange some more

$$\sum_{k=1}^{n} 4k^3 = (n+1)^4 - 1 - \sum_{k=1}^{n} 6k^2 - \sum_{k=1}^{n} 4k - n$$

Expand the term $(n+1)^4$ and pick up the $-1-n$:

$$\begin{gathered}(n+1)^4 - 1 - n \\ = n^4 + 4n^3 + 6n^2 + 4n + 1 - 1 - n \\ = n^4 + 4n^3 + 6n^2 + 3n\end{gathered}$$

Factor out an n

$$= (n)(n^3 + 4n^2 + 6n + 3)$$

And another $n+1$

$$= (n)(n+1)(n^2+3n+3)$$

Recall our previous results:

$$\sum_{k=1}^{n} 6k^2 = 6\sum_{k=1}^{n} k^2$$
$$= 6 \ \frac{n(n+1)(2n+1)}{6}$$
$$= n(n+1)(2n+1)$$

Similarly

$$\sum_{k=1}^{n} 4k = 4\sum_{k=1}^{n} k$$
$$= 4 \ \frac{n(n+1)}{2}$$
$$= 2n(n+1)$$

Substitute all three of these results (and pull out the factor of 4 from the sum):

$$4\sum_{k=1}^{n} k^3 = (n)(n+1)(n^2+3n+3) - n(n+1)(2n+1) - 2n(n+1)$$

Just a bit more algebra. See that we have $n(n+1)$ in each term. We have

$$= n(n+1) \ [\ (n^2+3n+3) - (2n+1) - 2 \ ]$$
$$= n(n+1) \ [\ n^2+3n+3-2n-1-2 \ ]$$
$$= n(n+1) \ [\ n^2+n \ ]$$
$$= n(n+1) \cdot n(n+1)$$

So all together we have

$$4\sum_{k=1}^{n} k^3 = n(n+1)\cdot n(n+1)$$

$$\sum_{k=1}^{n} k^3 = \frac{n(n+1)}{2}\cdot\frac{n(n+1)}{2}$$

$$\sum_{k=1}^{n} k^3 = \ [\ \frac{n(n+1)}{2}\ ]^2$$

A remarkable simplification!

Looking deeper

$$\sum_{k=1}^{n} k^3 = [\ \sum_{k=1}^{n} k\ ]^2$$

We want to try to understand something more about why this is true.

A web search revealed the answer. Here's an interesting pattern for the cubes of integers

$$1^3 = 1$$

$$2^3 = 8 = 3+5$$

$$3^3 = 27 = 7+9+11$$

$$4^3 = 64 = 13+15+17+19$$

$$5^3 = 125 = 21+23+25+27+29$$

If you want a formula for n^3, notice that the first term is $n^2 - n + 1$ and the last term is $n^2 - n + 2n - 1$, and the number of terms for each sum equals n. (There are n odd numbers between 1 and $2n - 1$).

In other words, the sum of all the cubes of integers from 1^3 to n^3 is equal to the sum of all the odd numbers up to $n^2 - n + 2n - 1 = n^2 + n - 1$.

How many of these numbers are there? A little thought should convince you that the answer is $(n^2+n)/2$. For example, with $n = 5$, our last odd number is $5^2+5-1 = 29$, and we have $(25+5)/2 = 15$ terms.

We want the sum of the first $(n^2 + n)/2$ odd numbers.

Let's look at another pattern

$$1 = 1$$

$$2^2 = 4 = 1 + 3$$

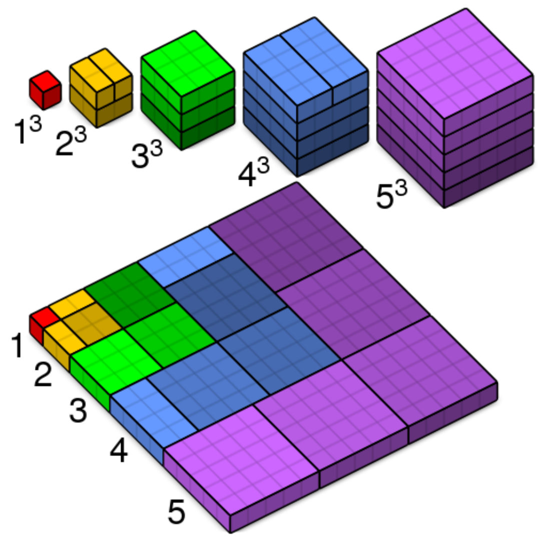
$$3^2 = 9 = 1 + 3 + 5$$

$$4^2 = 16 = 1 + 3 + 5 + 7$$

$$5^2 = 25 = 1 + 3 + 5 + 7 + 9$$

The *odd number theorem* says that the sum of the first n odd numbers is equal to n^2. We want the sum of the first $(n^2 + n)/2$ odd numbers, so that's $((n^2 + n)/2)^2$. And that's how we get our formula.

Here is another beautiful proof without words:



The length of the bottom pattern is a triangular number, which is itself a sum of squares. When squared it equals the sum of cubes.

another method

Here is another approach that I found very early in Hamming (Chapter 2) and have not seen in other books. It is called the method of *undetermined coefficients*.

We observe that the sum of integers formula has order n^2, while the sum of squares has order n^3, so we expect the sum of cubes would have n^4.

$$\sum_{k=0}^{k=n} k^3 = an^4 + bn^3 + cn^2 + dn + e$$

and if $n = 0$ the sum is zero so $e = 0$.

The right-hand side is

$$an^4 + bn^3 + cn^2 + dn$$

The inductive step is to write the formula for $m - 1$, and then add m^3 to it.

The right-hand side is just the formula, writing m for n

$$am^4 + bm^3 + cm^2 + dm$$

The left-hand side is the formula for $(m - 1)$, plus m^3 from the induction step:

$$a(m-1)^4 + b(m-1)^3 + c(m-1)^2 + d(m-1) + m^3$$

We work with the left-hand side. Expand each term using the binomial theorem:

$$a \; [\; m^4 - 4m^3 + 6m^2 - 4m + 1 \;]$$

$$b \; [\; m^3 - 3m^2 + 3m - 1 \;]$$

$$c \; [\; m^2 - 2m + 1 \;]$$

$$d \; [\; m - 1 \;]$$

Next, group the cofactors by the corresponding powers:

$$[\; a \;] \; m^4$$

$$[\; -4a + b + 1 \;] \; m^3$$

$$[\; 6a - 3b + c \;] \; m^2$$

$$[\; -4a + 3b - 2c + d \;] \; m$$

$$a - b + c - d$$

Now to the point. The cofactors for *each power* of m must cancel exactly.

am^4 cancels on left and right, likewise bm^3, cm^2 and dm. That leaves four equations.

$$-4a + 1 = 0$$

$$6a - 3b = 0$$

$$-4a + 3b - 2c = 0$$

$$a - b + c - d = 0$$

We find that $a = 1/4$, $b = 1/2$, $c = 1/4$, $d = 0$. So then finally the formula is

$$an^4 + bn^3 + cn^2 + dn$$

$$= \frac{n^4 + 2n^3 + n^2}{4}$$

$$\frac{(n^2 + n)^2}{2^2} = \left[\frac{n(n+1)}{2} \right]^2$$

which is exactly what we will have from other approaches.

Hamming uses this method to get a general formula, but we will not need that, because we will show how to use the binomial theorem to get what is necessary.

Part IX

Analytic geometry

Chapter 39

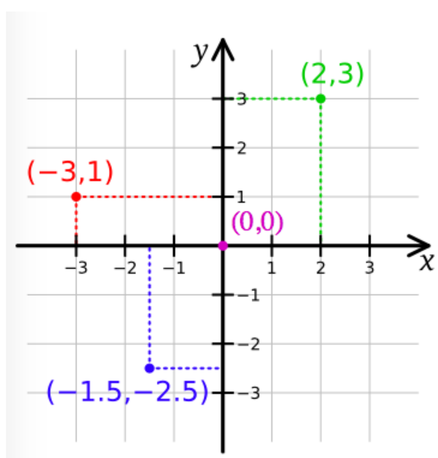
Lines and slopes

It is difficult today to put ourselves in the place of those who tried to reason about mathematics through the ages.

The Greeks lacked algebra, and although the Romans worked with numbers they did not have decimal notation. The concept of 0 came much later (from India), and even in the Middle Ages there was as yet no such thing as the equals sign =, which dates from 1557.

https://en.wikipedia.org/wiki/Table_of_mathematical_symbols_by_introduction_date

The invention of analytic geometry is often ascribed solely to Descartes, but Fermat also had his own version. There are two fundamental ideas.



The first is to orient two number lines on a piece of paper, at right angles, and then consider pairs of numbers (x, y) in the 2D plane. Such pairs or tuples are called points.

Descartes published this idea in 1637. The presentation would be difficult to recognize as our current system, but the germ is there: axes where the position of a variable could be marked. Only the positive numbers would be shown, and the axes not necessarily perpendicular. As to the proofs, here is wikipedia on the subject:

> His exposition style was far from clear, the material was not arranged in a systematic manner and he generally only gave indications of proofs, leaving many of the details to the reader. His attitude toward writing is indicated by statements such as "I did not undertake to say everything," or "It already wearies me to write so much about it," that occur frequently. In conclusion, Descartes justifies his omissions and obscurities with the remark that much was deliberately omitted "in order to give others the pleasure of discovering [it] for themselves."

The second idea of analytic geometry is to plot all the points that satisfy some mathematic relationship between x and y, for example the parabola $y = x^2$.

To do this, pick a few values of x and calculate the corresponding values of y. For example: $(0, 0), (\pm 1, 1), (\pm 2, 4), \ldots$. Plot these points, and then finally, sketch the graph of the curve, without actually trying to plot *all* of the individual points (of which there is an infinite number). We make the assumption here that the function being plotted is continuous, so that the sketch of a curve between two points that are close enough together will be fairly smooth and if the x-values are close to the plotted x, the corresponding y-values will not be not too different from the plotted y.

point

A point is simply an ordered pair (x, y) such as $(1, 3)$. Often points have integer components, but they don't have to be.

distance formula

The x- and y-axes are perpendicular to one another (a fancy word for that is *orthogonal*).

Suppose we pick two particular points (s, t) and (u, v), plot them on a graph, and then draw the line that connects them. Recall Euclid's first two postulates:

- A straight line segment can be drawn joining any two points.
- Any straight line segment can be extended indefinitely in a straight line.

The distance between the two points is given by the Pythagorean formula, where Δx is the change in x and Δy is the change in y:

$$d = \sqrt{\Delta x^2 + \Delta y^2}$$

It is often easier to use the squared distance and avoid the square root:

$$d^2 = \Delta x^2 + \Delta y^2$$

$$= (s - u)^2 + (t - v)^2$$

Switching the order of (s, t) and (u, v) doesn't change the result.

formulas for a line

Now we want to derive an equation that describes (is valid for) all the points or pairs of values (x, y) on this line. A general approach is to say that the line has some slope m, which is defined as Δy, divided Δx:

$$m = \frac{\Delta y}{\Delta x} = \frac{y - y'}{x - x'}$$

This is called the *point-slope equation.* For any two particular points (s, t) and (u, v) one can plot a line between them. The slope is

$$m = \frac{s - u}{t - v}$$

One can write the two points in either order, with the same result since:

$$\frac{s - u}{t - v} = \frac{u - s}{v - t}$$

Depending on the details, the value of m might be zero, for a horizontal line, where all the values of y are the same (which happens when $s = u$). Or it might be undefined, for a vertical line, where all the values of x are identical ($t = v$).

In most cases, however, $m \neq 0$ and $m \in (-\infty, \infty)$. That is, m is usually non-zero and not infinite.

Except in the case of the vertical line, we can write

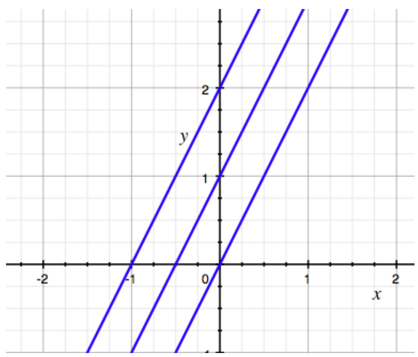
$$y = mx + y_0$$

for any point (x, y) on a given line, where y_0 is the y-intercept, the value of y obtained when $x = 0$.

[The choice of b for the y-intercept is the usual notation, but it conflicts with another b that we will see in a minute.]

$y = mx + y_0$ is the *slope-intercept equation* of the line.

The equation of a line is determined by both the slope and one point on the line, for example the y-intercept. One can draw a whole family of parallel lines with the same slope and different y-intercepts. Here are three lines $y = 2x + y_0$ for $y_0 = \{0, 1, 2\}$.



The value of x corresponding to $y = 0$ is the x intercept

$$x_0 = -\frac{y_0}{m}$$

The point-slope equation is easily derived from the second one. Suppose we have $y = mx + y_0$:

Plugging in for specific points (s, t) and (u, v) we have

$$t = ms + y_0$$

$$v = mu + y_0$$

Subtracting:

$$v - t = m(u - s)$$

which rearranges to give the desired result.

intersections

Often one has two lines (or curves) and we want to find the point(s) that lie on both. We might have

$$y = 2x - 1$$

$$y = -x + 8$$

Substitute from the second into the first:

$$2x - 1 = -x + 8$$

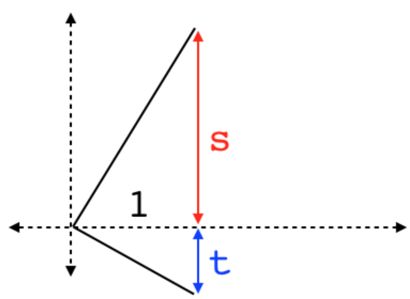
$$3x = 9$$

$$x = 3$$

From the first equation, $y = 5$, and we check that $x = 3, y = 5$ solves the second equation as well.

orthogonality

If two lines cross each other at right angles we say they are *orthogonal*. In that case the slopes have a special relationship. Their product is equal to -1.



Here is a simple proof. Draw two lines going through the origin, forming a right angle there. The first has slope s, so it goes through the point $(1, s)$, the second has slope -t and goes through $(1, -t)$.

From the Pythagorean theorem, the product of the two pieces of the base is equal to the altitude squared. Here:

$$st = 1^2 = 1$$

These are the lengths, i.e. the absolute values of the slopes.

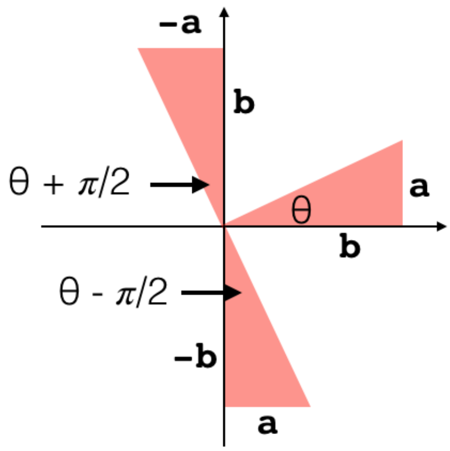
$$|s| \cdot \ |t| = 1$$

But clearly the sign of t is negative. So we arrive at

$$s \cdot (-t) = 1$$

$$m_1 \cdot m_2 = -1$$

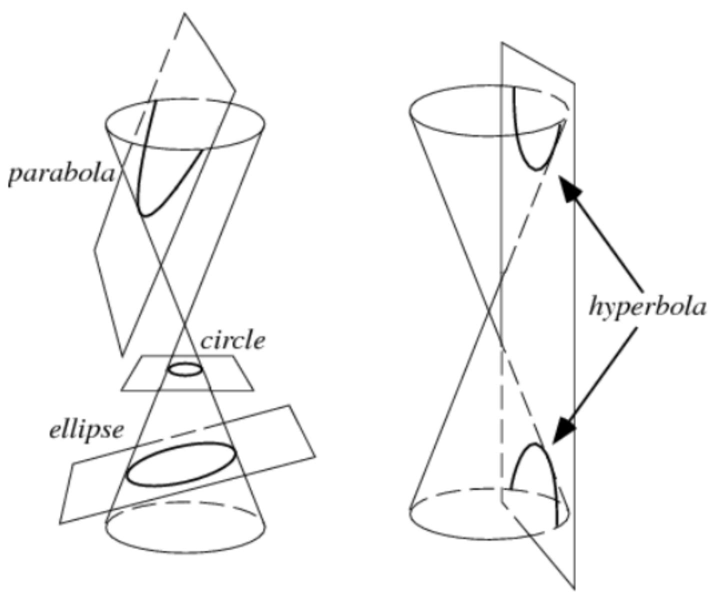
We'll see a natural easy proof of this once we look at trigonometry. Here is a hint:



Chapter 40

Circles again

We now consider what are called quadratic forms, as distinguished from linear equations (i.e., for lines). The quadratics contain a squared term (or a term that mixes x and y).



The simplest example is the equation for a unit circle centered at the origin:

$$x^2 + y^2 = 1$$

Pythagoras tells us that for a point (x, y), the square of the distance from the origin is $x^2 + y^2$. This equation describes all the points whose distance from the origin is equal to $\sqrt{1} = 1$. But all the points equi-distant to a point form a circle. We generalize

$$x^2 + y^2 = r^2$$

It is clear that when $y = 0$, $x = \pm\, r$, and when $x = 0$, $y = \pm\, r$. r is the radius of the circle.

Now, what happens if we displace the unit circle from the origin so its center is at $(1, 0)$? What this amounts to is adding 1 to the x value of every point. If we solve for x

$$x = \sqrt{1 - y^2}$$

and then add 1

$$\begin{aligned} x &= \sqrt{1 - y^2} + 1 \\ (x - 1)^2 &= 1 - y^2 \\ (x - 1)^2 + y^2 &= 1 \end{aligned}$$

Or, more generally

$$(x - h)^2 + (y - k)^2 = r^2$$

where the origin of the circle is at (h, k).

Multiplying out:

$$x^2 - 2hx + h^2 + y^2 - 2ky + k^2 = r^2$$
$$x^2 + y^2 - 2hx - 2ky + (h^2 + k^2 - r^2)$$

Comparing to the most general form for a quadratic

$$Ax^2 + By^2 + Cxy + Dx + Ey + F = 0$$

We see that

$$A = 1, \qquad B = 1, \qquad C = 0$$

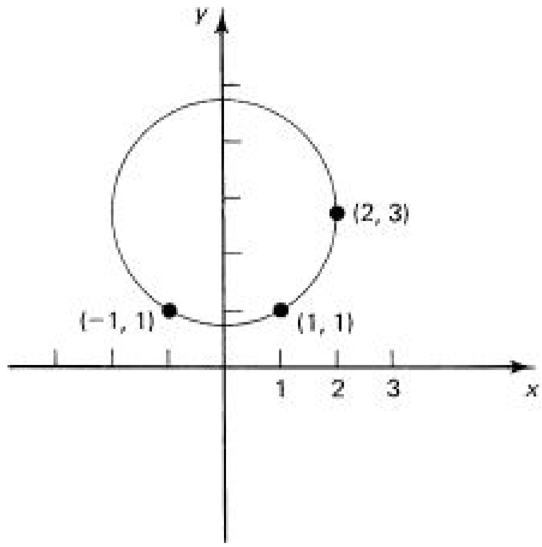
and in fact, this is true for all circles. (If $A = B \neq 1$, just divide all the terms by A).

Moreover

$$D = -2h, \qquad E = -2k, \qquad F = h^2 + k^2 - r^2$$

This equation can help us solve the following problem from Hamming: find the equation of the circle that passes through the following three points:

$$(-1, 1), (1, 1), (2, 3)$$



We write

$$x^2 + y^2 + Dx + Ey + F = 0$$

From the values of x and y at each of the three points we get

$$1 + 1 - D + E + F = 0$$

$$1 + 1 + D + E + F = 0$$

$$4 + 9 + 2D + 3E + F = 0$$

Three equations in three unknowns. We can do that.

Adding the first two equations together:

$$4 + 2(E + F) = 0$$

so $E + F = -2$.

Subtracting the first two equations (or substituting the result for $E+F$) tells us that $D = 0$.

Adding (-3) times the second equation to the third gives:

$$1 + 6 - D - 2F = 0$$

$$7 - 2F = 0$$

$F = 7/2$, and since $E + F = -2$, $E = -11/2$.

So the solution is

$$x^2 + y^2 - \frac{11}{2}y + \frac{7}{2} = 0$$

You can check that it works for all three points:

$$(-1,1),(1,1),(2,3)$$

The first two are easy, while the third gives

$$4 + 9 - \frac{11}{2}3 + \frac{7}{2} = 0$$

$$8 + 18 - 33 + 7 = 0$$

which looks correct.

completing the square

We can improve this by completing the square. We see that

$$y^2 - \frac{11}{2}y + (\frac{11}{4})^2 = (y - \frac{11}{4})^2$$

We must add that back to the right-hand side of the original to obtain:

$$x^2 + (y - \frac{11}{4})^2 = (\frac{11}{4})^2 - \frac{7}{2}$$

The center is at $(0, 11/4)$. The radius doesn't come out cleanly but r^2 is

$$\frac{121}{16} - \frac{56}{16} = \frac{65}{16}$$

so r is slightly more than 2.

Or recall that we had:

$$D = -2h, \quad E = -2k, \quad F = h^2 + k^2 - r^2$$

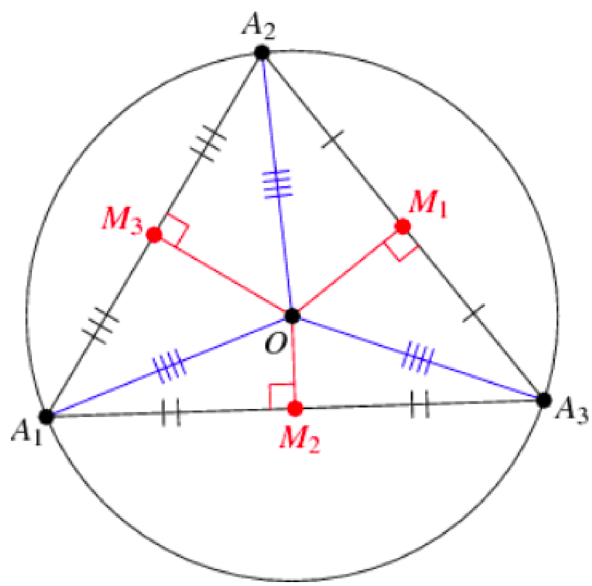
From this, we have that $h = 0$ and $k = -E/2 = 11/4$, and the radius is more complicated, as we said.

plane geometry

We can check our work by solving the problem using a technique from plane geometry. Again, we want the circle passing through three points:

$$(-1,1),(1,1),(2,3)$$

Take two of the points to be placed on a circle and construct the line segment joining them (a chord of the circle). Find the midpoint of the chord and erect a perpendicular bisector through the midpoint. Now, every point lying on the bisector is equidistant from the two starting points. Proof: draw the two triangles including that point, the two starting points and the midpoint of the bisector. The two triangles are congruent. Here is the general picture.

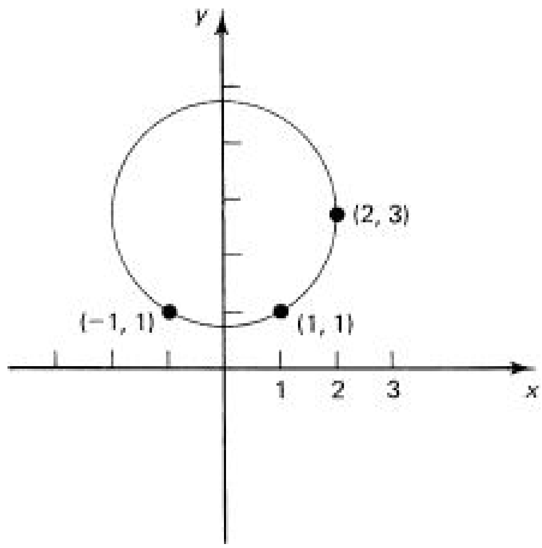


It's a bit trickier to prove that *every* point that is equidistant from the two points lies on the bisector. We assume that.

Since every point that is equidistant from the two points lies on the bisector, the radius of the circle lies on the bisector.

Then, erect a perpendicular bisector of a chord joining another pair chosen from the three points. This new bisector and the first one meet at the center of the circle.

In our case two points $(-1,1),(1,1)$ are symmetric about the y-axis. Therefore it is clear that the perpendicular bisector for these two points is the y-axis.



For the second bisector, form the vector between $(1, 1)$ and $(2, 3)$ as $\mathbf{v} = \langle 1, 2 \rangle$. The midpoint is at $(1, 1) + \mathbf{v}/2 = (3/2, 2)$.

The slope of the bisector is the negative inverse of the slope for the chord which is $-1/2$ so the equation of the bisector is

$$y - y_0 = -\frac{1}{2}(x - x_0)$$

Plugging in the point that we know, we obtain

$$y - 2 = -\frac{1}{2}(x - 3/2)$$

We want to solve for y when $x = 0$, crossing the first bisector, the y-axis

$$y - 2 = -\frac{1}{2}(-3/2)$$

$$y = \frac{11}{4}$$

So the center is at $(0, 11/4)$, which matches what we had before. We compute the distance to one of the points $(1, 1)$ as

$$d = \sqrt{1^2 + (11/4 - 1)^2} = \sqrt{1 + 49/16}$$

which also matches our previous result.

quadratics

The technique of completing the square comes from the standard equation

$$(x+p)^2 = x^2 + 2px + p^2$$

We run into problems where we have the $2px$ but not the p^2. For example

$$x^2 + y^2 + Dx + F = 0$$

Focus on

$$x^2 + Dx$$

we want to turn this into

$$(x + \text{something})^2$$

if D is like $2p$ we need to add something like p^2:

$$x + Dx + \frac{D^2}{4}$$

$$= (x + \frac{D}{2})^2$$

Since we added $D^2/4$ on the left, we must also add it on the right. We obtain

$$(x + \frac{D}{2})^2 + y^2 + F = \frac{D^2}{4}$$

The general equation for a quadratic is

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

In starting to work with one of these, the first thing to do is to see if there is a term which "mixes" x and y, that is, whether there is some term like Bxy. If there is, we might think about rotating the curve so that it is in a standard orientation.

Let us assume we've done that, we relabel the new A, C etc. and assume here that $B = 0$.

Once in standard orientation, the next thing we might do is to translate the quadratic so that it is centered on the origin. We do that by completing the square for both x and y. After that, consider cases:

• Both A and C are non-zero, B = 0, and $F < 0$. If

∘ A and C are both > 0: it's an ellipse.

∘ A and C are of opposite signs: it's a hyperbola.

∘ A, C and F are all negative: it's imaginary.

• Only one squared term is present, but we still have the other variable

$Ax^2 + Ey + F = 0$: it's a hyperbola.

Not every quadratic equation gives a conic. Some are "degenerate". For example, having done all the right manipulations, we might end up with something like

$$A'(x-h)^2 + B'(y-k)^2 = 0$$

which has only $x = h$ and $y = k$ as a solution. It's a point.

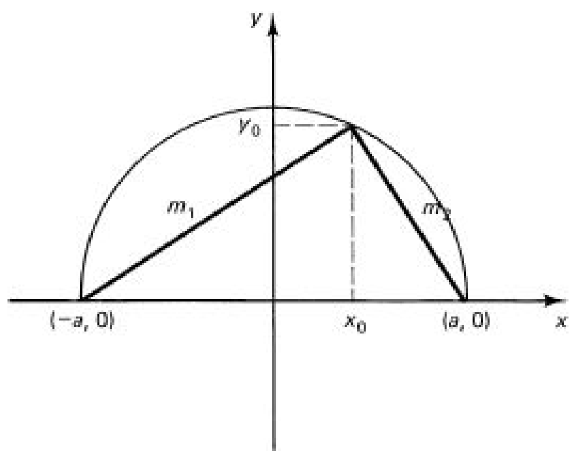


Figure 6.2-3 Angle in a semicircle

Here is another problem from Hamming. We need to prove that the angle above is a right angle. We know this is true from geometry. But now we wish to practice are analytic geometry.

Suppose the equation of the circle is

$$x^2 + y^2 = a^2$$

The point on the circle is (x_0, y_0).

Our first solution uses slopes and points. The line from $(-a, 0)$ to (x_0, y_0) has slope

$$m_1 = \frac{y_0}{x_0 + a}$$

The line from $(a, 0)$ to (x_0, y_0) has slope

$$m_2 = \frac{y_0}{a - x_0}$$

Two lines meet at a right angle if the product of their slopes is equal to -1.

$$m_1 m_2 = \frac{y_0}{x_0 + a} \ \frac{y_0}{a - x_0}$$

$$= \frac{y_0^2}{a^2 - x_0^2} = \frac{y_0^2}{x_0^2 + y_0^2 - x_0^2} = -1$$

This was not pretty, it's just good exercise.

And here is a proof using vectors and the dot product. Consider the semicircle centered on the origin with radius a, so the ends of the diameter are at $(x = \pm\, a, 0)$.

Form the vectors from those ends to an arbitrary point (x, y) on the perimeter:

$$\mathbf{u} = \ \langle x + a, y \rangle$$

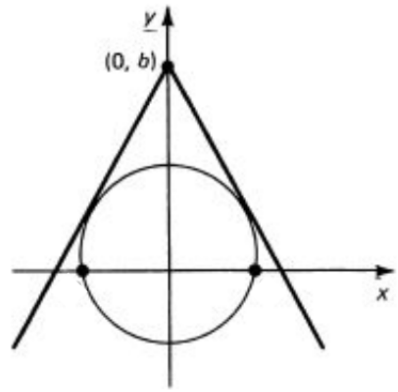
$$\mathbf{v} = \ \langle x - a, y \rangle$$

Notice that

$$\mathbf{u} \cdot \mathbf{v} = (x + a)(x - a) + y^2$$
$$= x^2 - a^2 + y^2 = 0$$

because $x^2 + y^2 = a^2$ for any point on the circle.

As our last example, consider the problem of finding the equation of a line tangent to a circle that goes through some arbitrary point b.



We take the circle to have radius a and be centered at the origin. We take the point b to be on the y-axis. The equation of the line on the right side is

$$\frac{y - y_0}{x - x_0} = m = \frac{y - b}{x}$$

$$y = mx + b$$

(well, of course).

For the point or points where the line intersects the circle we also have

$$y = \sqrt{a^2 - x^2}$$

$$\sqrt{a^2 - x^2} = mx + b$$

$$a^2 - x^2 = m^2x^2 + 2bmx + b^2$$

$$(m^2 + 1)x^2 + 2bmx + b^2 - a^2 = 0$$

From the quadratic equation:

$$x = \frac{-2bm \pm \sqrt{4b^2m^2 - 4(m^2 + 1)(b^2 - a^2)}}{2(m^2 + 1)}$$

We are looking for the case where there is a single solution so the discriminant under the square root must be equal to zero:

$$4b^2m^2 = 4(m^2 + 1)(b^2 - a^2)$$

$$m^2b^2 = m^2b^2 - m^2a^2 + b^2 - a^2$$

$$0 = -m^2a^2 + b^2 - a^2$$

$$m = \pm\frac{\sqrt{b^2 - a^2}}{a}$$

This makes sense since if $a = b$ the single tangent should be horizontal with zero slope. Notice that if $a^2 > b^2$ there is no real solution. This corresponds to having b inside the circle.

Part X

Trigonometry

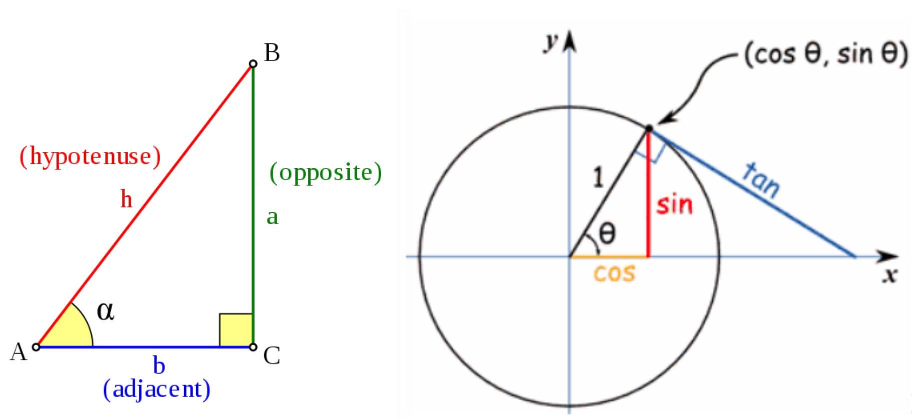
Chapter 41

Six functions

basic definitions

The most elementary trigonometric functions are the sine and cosine. These are defined in geometry as ratios of the lengths of the sides of a right triangle.

Looking at the left panel, we say that the sine of the angle α is the ratio *opposite-over-hypotenuse*, while the cosine of α is the ratio *adjacent-over-hypotenuse*. Tangent is the ratio *opposite-over-adjacent*. The names are abbreviated to three letters in formulas.



Using the notation for the sides from the figure:

$$\sin\alpha = \frac{a}{h}, \qquad \cos\alpha = \frac{b}{h}, \qquad \tan\alpha = \frac{a}{b} = \frac{\sin\alpha}{\cos\alpha}$$

The "unit circle" is a circle of radius 1 with its center positioned at the origin of coordinates, the place where the x and y axes cross. From the right panel of the diagram you can see that any point (x, y) on the unit circle can be described in radial coordinates as

$$x = \cos\theta \qquad y = \sin\theta$$

In the diagram, all three right triangles are similar because the red line is an altitude of the largest right triangle. Thus, by similar triangles, the blue side has this relationship

$$\frac{\text{blue side}}{1} = \frac{\sin\theta}{\cos\theta}$$

which explains why it is labeled as tan (α).

If the vertex labeled B is denoted angle β (the complementary angle of α), then the notions of opposite and adjacent switch so that:

$$\sin\alpha = \cos\beta, \qquad \cos\alpha = \sin\beta$$

If the circle has radius r then

$$x = r\cos\theta \qquad y = r\sin\theta$$

Stewart:

> The mathematicians of ancient India built on the Greek work to make major advances in trigonometry. They [used] the sine (sin) and cosine (cos) functions, which we still do today. Sines first appeared in the Surya Siddhanta, a series of Hindu astronomy texts from about the year 400, and were developed by Aryabhata in Aryabhatiya around 500. Similar ideas evolved independently in China.

The other functions are the inverses of sine, cosine and tangent, namely: cosecant, secant and cotangent. The secant (inverse cosine) comes up somtimes, but the other two are not especially important in calculus.

However, there is one context that we will look at, namely, Archimedes determination of the value of π. The crucial step in that approach will turn out to be the calculation of the cotangent of the half-angle $\theta/2$ given the values of cotangent and cosecant for angle θ.

The main relationship or identity is derived from the Pythagorean theorem. We had above that for a unit circle

$$x = r\cos\theta \qquad y = r\sin\theta$$

Since x and y are the sides of a right triangle whose hypotenuse is r

$$x^2 + y^2 = r^2$$

and for a unit circle

$$\cos^2\theta + \sin^2\theta = 1$$

which is usually written

$$\sin^2\theta + \cos^2\theta = 1$$

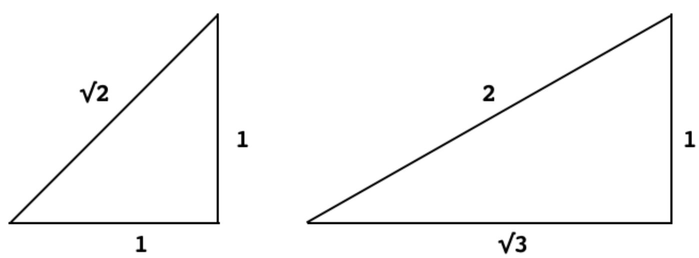
and transformed to

$$1 + \tan^2\theta = \sec^2\theta$$

particular values

We can easily determine the values for these functions for three special cases.

The first is the angle 45 degrees or $\pi/4$. Draw an isosceles right triangle with sides of length 1 (left panel).



Then the hypotenuse has length $\sqrt{2}$ (from Pythagoras) and the values are

$$\sin\frac{\pi}{4} = \frac{1}{\sqrt{2}} = \cos\frac{\pi}{4}$$

$$\tan\frac{\pi}{4} = 1$$

For the other two, bisect an equilateral triangle and erase one half (right panel). The smaller angle is 30 degrees or $\pi/6$ and its complement is 60 degrees or $\pi/3$.

The values are

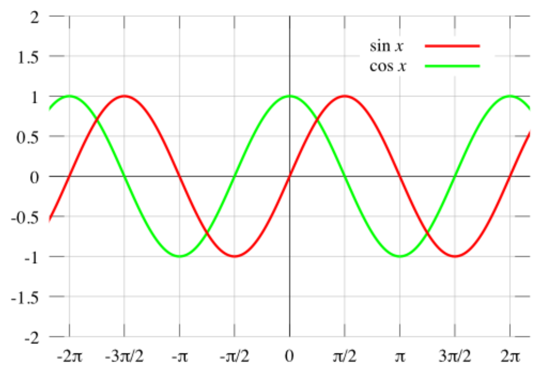
$$\sin\frac{\pi}{6} = \frac{1}{2} = \cos\frac{\pi}{3}, \qquad \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2} = \sin\frac{\pi}{3}$$

$$\tan\frac{\pi}{6} = \frac{1}{\sqrt{3}}$$

We can easily verify that

$$(\frac{1}{2})^2 + (\frac{\sqrt{3}}{2})^2 = 1, \qquad (\frac{1}{\sqrt{2}})^2 + (\frac{1}{\sqrt{2}})^2 = 1$$

graph



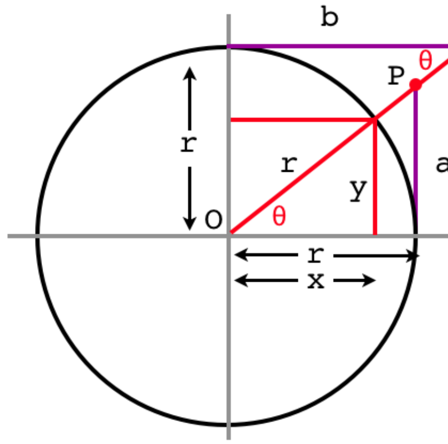
Savov:

> The sine function represents a fundamental unit of vibration. The graph of $\sin(x)$ oscillates up and down and crosses the x-axis multiple times. The shape of the graph of $\sin(x)$ corresponds to the shape of a vibrating string.

Imagine a circle placed to the left of a graph. I think of the sine function as the "shadow" of the point (x, y) as it travels around the circle at the same constant speed as the point on the graph "moves" to the right.

visualization of all six functions

Consider a unit circle. Extend the radius with the angle θ and then draw the vertical and horizontal tangents to the circle a and b.



The original triangle with sides x, y, r is similar to the triangle with sides r, a, OP, and both are similar to the triangle with sides b, r, OQ.

$$x, y, r \sim r, a, OP, \sim b, r, OQ$$

By similar $\triangle$

$$\frac{a}{r} = \frac{y}{x} = \tan\theta$$

But $r = 1$ so

$$a = \tan\theta$$

If you imagine a point moving around the circle a will get very large as $\theta \to \pi/2$, and in fact, approaches ∞ there (becomes undefined).

The segment OP is (by similar $\triangle$) to r as

$$\frac{OP}{r} = \frac{r}{x}$$

$$OP = \frac{1}{\cos\theta} = \sec\theta$$

The horizontal from the y-axis to Q is b. Consider θ near the top of the figure. By similar $\triangle$, the relations we had were

$$r/b = y/x = \tan\theta$$

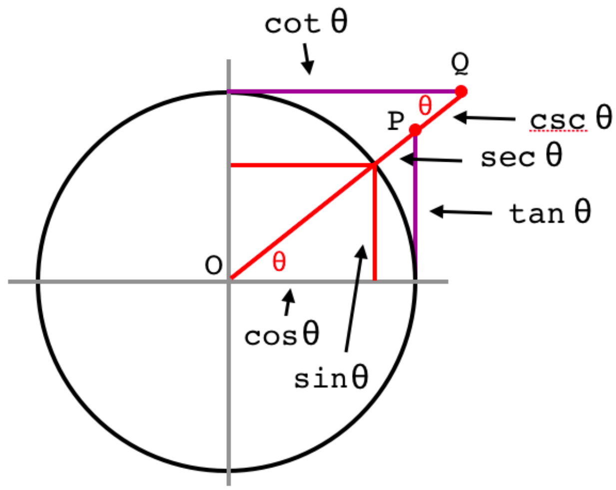
since $r = 1$

$$b = \frac{r}{\tan\theta} = \frac{1}{\tan\theta} = \cot\theta$$

Finally

$$r/OQ = 1/OQ = \sin\theta$$

$$OQ = \frac{1}{\sin\theta} = \csc\theta$$



Chapter 42

Sum of angles

cosine of a sum

The sum of angle formulas (i.e. formulas for the sine and cosine of the sum or difference of two angles) are used often in calculus, not only for working problems, but even in finding an expression for the "derivative" of sine and cosine.

You really must know them. I think it's so important that we will show three ways of finding these formulas — not all in this chapter. The easiest way to remember them uses Euler's equation, and we won't be ready for that until later. See **here**.

There are four equations: $\sin s \pm t$ and $\cos s \pm t$.

I've memorized only this one:

$$\cos s - t = \cos s \cos t + \sin s \sin t$$

By $\cos s - t$ we mean $\cos(s - t)$, but have left off the parentheses.

Say "cos cos" and then recall the difference in sign.

check

I like this version because it can be checked easily. Set $s = t$:

$$\cos s - t = \cos 0 = 1 = \cos^2 s + \sin^2 s$$

which is our favorite trigonometric identity and obviously correct.

change signs

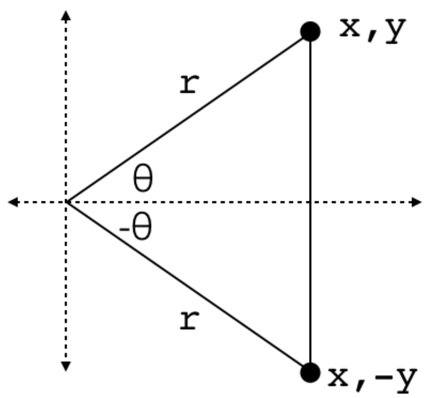
For $\cos s + t$ flip the sign on the second term.

$$\cos s + t = \cos s \cos t - \sin s \sin t$$

This is simply a result of the fact that

$$\cos -\theta = \cos \theta$$

$$\sin -\theta = -\sin \theta$$



The diagram shows the reason: $\cos \theta = \cos -\theta = x/r$ while $\sin \theta = y/r = -(\sin -\theta) = -(-y/r)$.

Proof:

$$\cos(s - (-u)) = \cos s \cos(-u) + \sin s \sin(-u)$$

Since $\cos -x = \cos x$ and $\sin -x = -\sin x$:

$$\cos(s + u) = \cos s \cos u - \sin s \sin u$$

But u is just a dummy variable (it could be any symbol), so

$$\cos(s + t) = \cos s \cos t - \sin s \sin t$$

sine of a sum

We will look at the proof for the sine formula later, for now just write it:

$$\sin s + t = \sin s \cos t + \sin t \cos s$$

Say "sin cos" and then, that here + goes with +. Like most things having to do with sine and cosine, there is a change of sign when moving from one to the other.

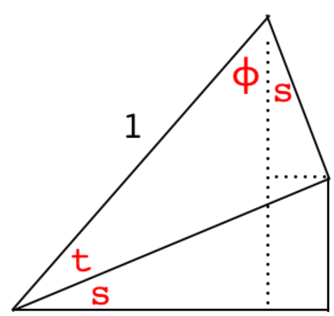
For $\sin s - t$, flip the sign on the second term, as before.

proof

Here is a geometric proof of both of the sum of angles formulas, using similar triangles. The key is to draw an inspired diagram.

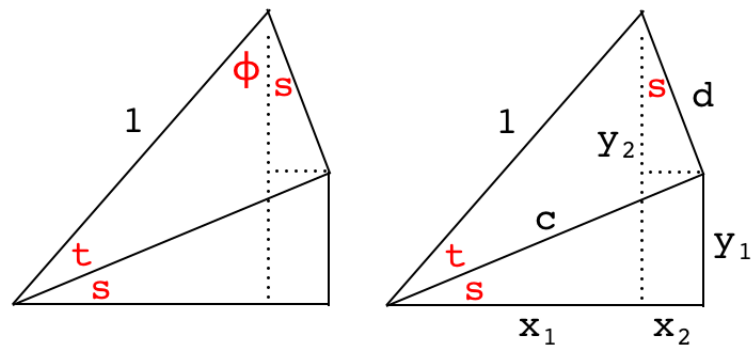
Consider a right triangle, with one of the angles labeled s. Construct another right triangle containing angle t, and scale it so that the base adjacent to angle t is just as long as the hypotenuse of the triangle containing angle s, and draw them one on top of the other as shown:

Scale the joined triangles so that the hypotenuse of the second triangle has unit length. Our crucial insight is to draw vertical and horizontal dotted lines as shown below.



The angle s is part of a right triangle with angle t adjacent, where the third acute angle is ϕ. But ϕ is also part of a second right triangle containing t plus the angle adjacent to ϕ. Therefore, that adjacent angle is also equal to angle s.

We add some labels to the sides of the triangles and calculate the sine and cosine of s, t and $s + t$:



Since I already know the result I am looking for, I write what we had before

$$\cos s \cos t - \sin s \sin t$$

From the figure

$$\cos s = \frac{x_1 + x_2}{c}; \quad \cos t = \frac{c}{1}; \quad \cos s \cos t = x_1 + x_2$$

The sine of s is a little trickier, look at the small right triangle at the top of the figure

$$\sin s = \frac{x_2}{d}; \quad \sin t = \frac{d}{1}; \quad \sin s \sin t = x_2$$

The difference is

$$\cos s \cos t - \sin s \sin t = x_1$$

but from the diagram it's clear that

$$\cos s + t = x_1$$

□

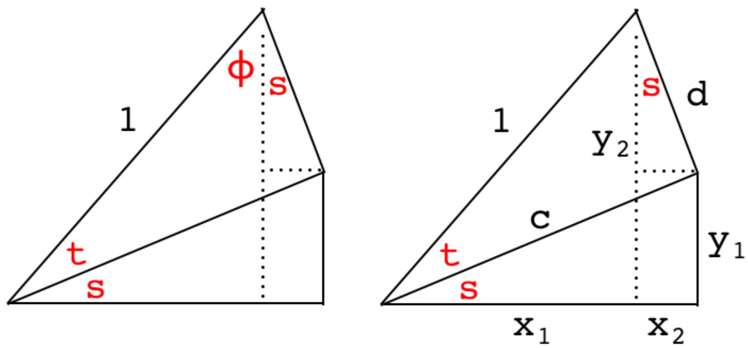
As a quick check we can ask what happens to the formula

$$\cos s + t = \cos s \cos t - \sin s \sin t$$

when $t = 0$. Then the first term is the cosine of s, and the second term is equal to 0. The formula is symmetrical with respect to s and t.

extension to sine

Referring back to the diagram (and again, with our goal clearly in mind)



$$\sin s = \frac{y_1}{c}; \quad \cos t = \frac{c}{1}; \quad \sin s \cos t = y_1$$

$$\sin t = \frac{d}{1}; \quad \cos s = \frac{y_2}{d}; \quad \sin t \cos s = y_2$$

But

$$\sin s + t = y_1 + y_2 = \sin s \cos t + \sin t \cos s$$

Using the even/odd function rules, we get

$$\sin s - t = c + d = \sin s \cos t - \sin t \cos s$$

And that's all four of them.

double-angle formulas

Very quickly, sine:

$$\sin s + t = \sin s \cos t + \cos s \sin t$$

$$\sin 2t = 2 \sin t \cos t$$

And cosine:

$$\cos s + t = \cos s \cos t - \sin s \sin t$$

$$\cos 2t = \cos^2 t - \sin^2 t = 2 \cos^2 t - 1$$

another calculation

We found previously that

$$\sin\frac{\pi}{4} = \cos\frac{\pi}{4} = \frac{1}{\sqrt{2}}$$

$$\sin\frac{\pi}{6} = \cos\frac{\pi}{3} = \frac{1}{2}; \qquad \sin\frac{\pi}{3} = \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2}$$

These angles correspond to 30, 45 and 60 degrees. It might be nice to have sine and cosine of 15 and 75 degrees as well. That would make even divisions of the first 90 degrees. We can get them as the sum and difference of $\pi/4$ and $\pi/6$.

Let $s = \pi/4$ and $t = \pi/6$. Then

$$\begin{aligned}
\sin\frac{\pi}{12} &= \sin s - t = \sin s \cos t - \sin t \cos s \\
&= \frac{1}{\sqrt{2}} \cdot \frac{\sqrt{3}}{2} - \frac{1}{2} \cdot \frac{1}{\sqrt{2}} = \frac{\sqrt{3}-1}{2\sqrt{2}} \\
\cos\frac{\pi}{12} &= \cos s - t = \cos s \cos t + \sin s \sin t \\
&= \frac{\sqrt{3}}{2} \cdot \frac{1}{\sqrt{2}} - \frac{1}{2} \cdot \frac{1}{\sqrt{2}} = \frac{\sqrt{3}+1}{2\sqrt{2}}
\end{aligned}$$

We just check that $\sin^2\theta + \cos^2\theta = 1$:

$$\frac{(\sqrt{3}-1)^2 + (\sqrt{3}+1)^2}{(2\sqrt{2})^2}$$

$$= \frac{3 - 2\sqrt{3} + 1 + 3 + 2\sqrt{3} + 1}{8} = 1$$

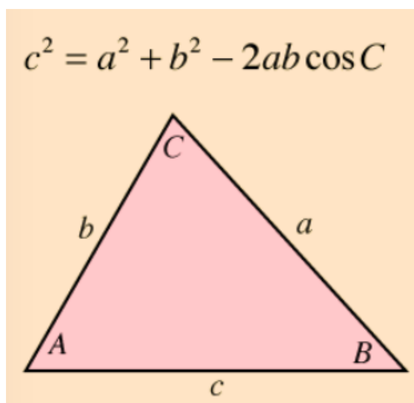
We can calculate similarly for $s+t = 5\pi/12$ or just switch sine and cosine from $\pi/12$.

Chapter 43

Law of cosines

Law of cosines

Designate the lengths of a triangle's sides as a, b, c and the angle between sides a and b as C (because it is opposite side c). The law of cosines says that



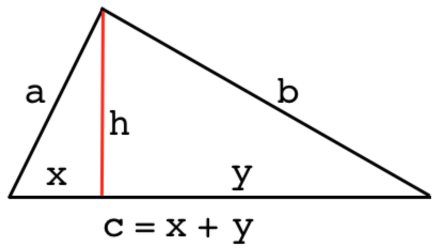
$$c^2 = a^2 + b^2 - 2ab\cos C$$

Lockhart calls this the "generalized" Pythagorean theorem. We can view the term $-2ab\cos C$ as a correction term which disappears in the case where $\angle C$ is 90 degrees.

derivation

The result follows from the Pythagorean Theorem. (In fact, we can reuse the same diagram that was shown for the algebraic proof of the theorem).

For a triangle with sides a, b and c and angles opposite those sides A, B and C, divide the third side into two lengths $c = d + e$ using the vertical altitude from vertex C.



$$a^2 - x^2 = h^2$$
$$b^2 - y^2 = h^2$$

So

$$a^2 = x^2 + h^2 = x^2 + b^2 - y^2$$

But

$$y = c - x$$
$$y^2 = c^2 - 2cx + x^2$$

Therefore:

$$a^2 = x^2 + b^2 - (c^2 - 2cx + x^2)$$
$$a^2 = b^2 - c^2 + 2cx$$

Rearranging

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

Finally, $x = a\cos B$ so

$$a^2 = b^2 - c^2 + 2ac\cos B$$

This is the law of cosines.

□

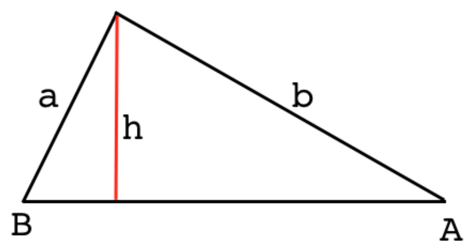
Any side of a triangle can be expressed in terms of the other two and the cosine of the angle between them. Thus, for example

$$c^2 = a^2 + b^2 - 2ab\cos C$$

$$a^2 = b^2 + c^2 - 2bc\cos A$$

Law of sines

I'll just mention another identity called the law of sines. In contrast to the law of cosines, it is fairly trivial.



$$\frac{h}{b} = \sin A$$

$$\frac{h}{a} = \sin B$$

Therefore

$$h = b\sin A = a\sin B$$

$$\frac{\sin A}{a} = \frac{\sin B}{b}$$

We could do the same construction and argument with either A or B, and the third angle, call it C. Therefore

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}$$

The sine of an angle, divided by the length of the side opposite, is a constant.

Part XI

Two basic operations in calculus

Chapter 44

Simple slopes

To introduce the two fundamental ideas in calculus, consider two measuring devices used while driving a car. Most good drivers look fairly often at the speedometer, which measures speed or velocity, or how fast you're going.

On the other hand, if someone gives you directions like — go three and a half miles and then turn left (where the old gas station used to be), you need to be watching your odometer.

Distance divided by time is velocity. Velocity times time equals distance. We can think of speed and velocity as the same for now.

Velocity is the *rate of change* of distance with time, it has units like miles per hour or feet per second (15 mph is exactly 22 feet per second; $15 \cdot 5280 = 22 \cdot 3600$).

In calculus we say that

- velocity is the **derivative** of the distance with respect to time
- distance is the **integral** of the velocity with respect to time

We can speak of velocity at a particular time t, as in "our current velocity is 60 miles per hour." But the distance, the integral, must be evaluated between appropriate starting and stopping points for the time.

In our example, you must first look at your odometer *before* you start on that 3.5 mile drive, and subtract the initial from the final value.

time-dependence

Distance equals velocity times time.

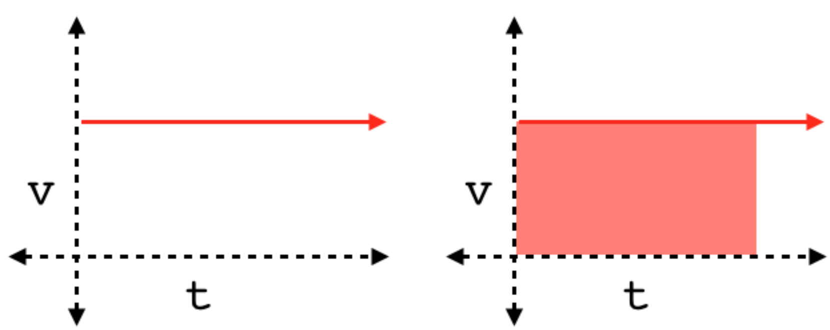
This is easy if the velocity is constant. Travel west on the interstate at exactly 60 miles per hour for 2 hours and your distance will be 120 miles from where you started (provided you don't start in Los Angeles).

It is standard to use s to refer to the distance traveled and v for velocity. If the velocity is constant then:

$$s = vt$$

According to the internet, s is from the Latin "spatium", for "space, room, or distance."

Suppose we plot velocity as a *function of time* with v on the y-axis and t on the x-axis.



Since the velocity is constant, the result is a straight horizontal line.

Furthermore, the distance traveled is the *area under the curve* (and above the x-axis) which is the area of a rectangle with sides v and t and as we said

$$s = vt$$

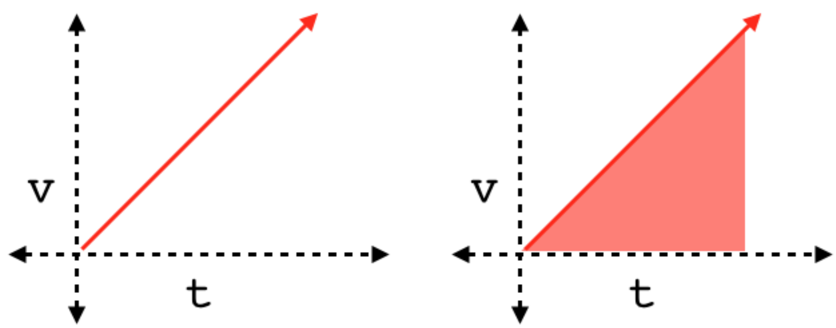
However, for most interesting problems the velocity is not constant.

Imagine maintaining pressure on the gas pedal in the car steadily so that, starting from a stop at zero time, after 1 second your velocity is 10 mph, after 2 seconds it is 20 mph, after 3 seconds, 30 mph. If we continue at the same rate of acceleration, we'll go from 0 to 60 mph in 6 seconds, which is quite a respectable time.

This example has constant acceleration. Here, we say that v is a constant function of time, and write

$$v = at$$

where a is the acceleration.



What about the distance? It turns out that the distance is once again the area under the curve.

Since a isn't zero, v must change with time.

If a is non-zero and constant, then v changes at a constant rate. Starting from 0, the final velocity will be $v = at$, but the distance traveled is no longer the product

$$s = v \times t \stackrel{?}{=}$$

because this v is the final velocity and that is not the correct v to use. For variable velocity, the distance traveled is the *average* velocity times the time. For smooth (constant) acceleration from zero to v, the average velocity is the average of the initial and final velocities:

$$v_{\text{avg}} = \frac{1}{2}\,(v_i + v_f) = \frac{1}{2}\,v$$

So the correct equation is:

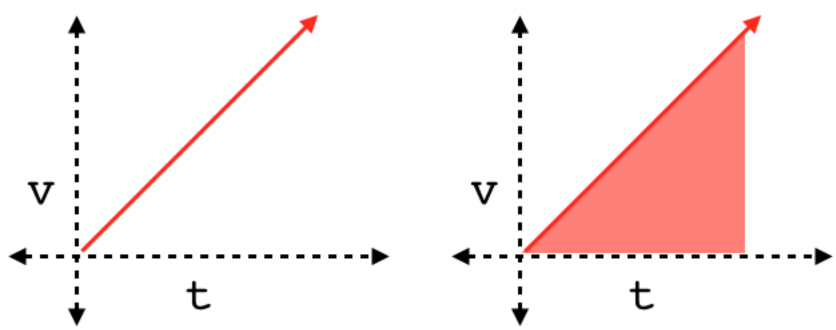
$$s = v_{\text{avg}}\, t = \frac{1}{2}\, v \cdot t$$

and since $v = at$

$$s = \frac{1}{2} a t^2$$

In this case, if we plot velocity as a function of time, we obtain a straight line that extends diagonally up with respect to the x-axis. The distance traveled is the area under the curve, below the line and above the x-axis.

The shape whose area is needed is a triangle. This also accounts for the factor of 1/2.



You probably know that if a mass m is dropped from a tall building like the Tower of Pisa, then the distance it has fallen goes like the square of the time. The equation is:

$$s = \frac{1}{2} g t^2$$

where g is the acceleration due to gravity.

Notice that this is the same equation as we just obtained.

The reason is that g is approximately constant near the surface of the earth, its value is about 10 in units of m/s^2. A fall of four seconds is about 80 meters.

Galileo knew this formula (at least, he knew the t^2 part of it), which he obtained not from experiments at the Tower of Pisa, but by timing the descent of balls down an inclined plane.

initial position and velocity

If you want to be more complete and say that the starting point is not necessarily the origin of the coordinate system, add a constant s_0 to describe the initial distance from the origin and obtain:

$$s = vt + s_0$$

and similarly, a constant v_0 to describe the initial velocity as shown above.

The full equation of motion is

$$s = \frac{1}{2}at^2 + v_0 t + s_0$$

We'll say much more about this later.

power rule

We will introduce the theory of calculus more formally in the next section of the book. For now, we just talk about a simple rule called the power rule.

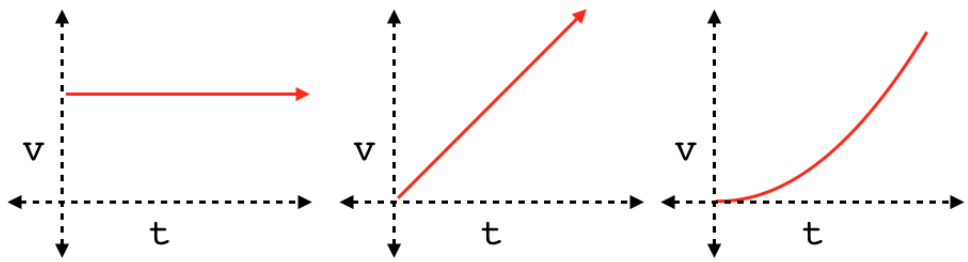
Switching notation to y and x, suppose that y is a *function* of x and write $y = f(x)$.

Here are three types of dependency (with c as a constant), with three corresponding types of graph.

- $y = c$

- $y = cx$
- $y = cx^2$

These are (respectively) the equations of: (i) a horizontal line, since y is constant, (ii) any other non-vertical line (y is proportional to x), and (iii), a parabola.



Suppose we are at some particular point on the curve, x.

We ask "what happens if we change x a little bit" and use the notation dx to refer to this little bit of x.

What happens to y? y will usually change by a small amount. Call that amount dy.

case 0

We can call this case 0 because we can write it as

$$y = cx^0 = c$$

Of course, in this case

$$y = c$$

y does not actually depend on x at all. The change dy resulting from a change in x, dx, is zero. That is what the curve plotted above tells us (left panel).

$$y = c, \qquad dy = 0 \cdot dx$$

The ratio dy/dx is the slope of the curve formed by plotting y against x. We call that slope the *derivative* of the function $f(x)$.

Divide both sides by dx and rewrite the above as:

$$\frac{dy}{dx} = 0$$

This plot is a horizontal line with slope 0.

case 1

Here, y is a linear function of x, the change dy is the change dx multiplied by c:

$$y = cx, \qquad dy = c \cdot dx$$

rearranging.

$$\frac{dy}{dx} = c$$

In analytical geometry, we calculate the slope of a line as $\Delta y / \Delta x$.

For a line, the slope is constant and so it doesn't matter which two points with coordinates $(x, y), (x', y')$ we choose for the calculation. The following is true for *any* two points on the line:

$$m = \frac{\Delta y}{\Delta x} = \frac{y - y'}{x - x'}$$

Above we had the example where $v = at$ with constant a. Then $dv/dt = a$.

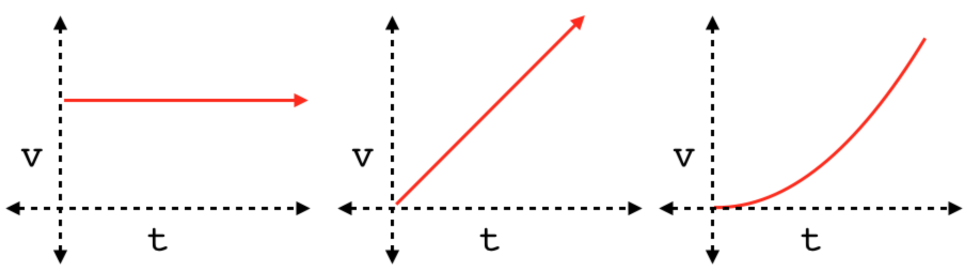
case 2

This case is different.

$$y = cx^2$$

We finally get to using some calculus.

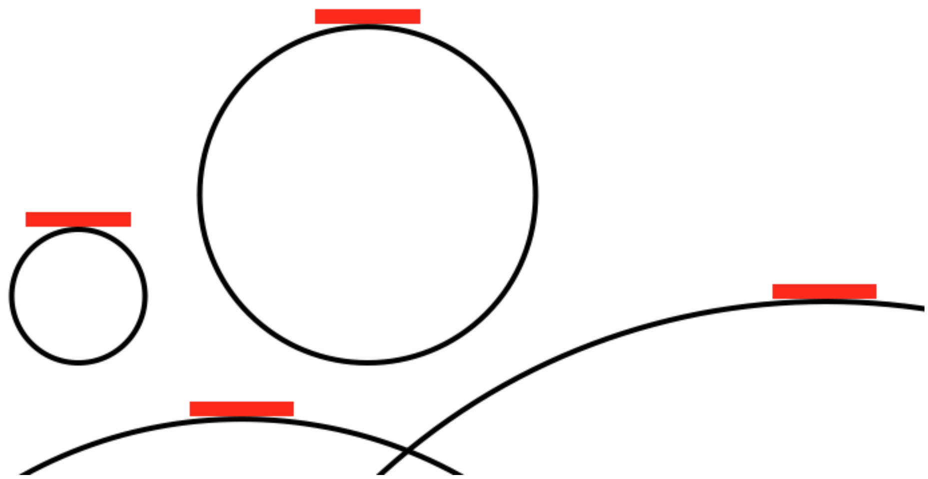
For a parabola, the slope of the curve at a point (the slope of the tangent to the curve $y = cx^2$) depends on the choice of x. The slope is steeper the further out you go in a positive direction on the x-axis (right panel).



It seems impossible to compute the slope of this curve in the standard way, by picking two points (x, y) and (x', y') and then calculating $\Delta y / \Delta x$, because the slope changes as we go out along the curve. You'll get a different answer for each different x.

key idea

The insight is that if x is sufficiently close to x' the slope is approximately constant. It's like saying that the earth is flat *locally*. If you detect any curvature, just zoom in a bit. In the figure below



the distance to the circle from the end of a line of fixed length decreases as we increase the size of the circle.

In calculus, we don't make the curves larger, we make the distance between x and x' smaller and smaller until it is so small, that the circle, or the parabola in the other figure, becomes flat. I have just magnified the figure so we could see it.

The error, the distance between the end of the red line above and the circle, gets smaller and smaller as a fraction of the line's length. Our approximation gets better and better. Just zoom in until the line is a good enough approximation to the shape of the circle, if the curve doesn't look flat enough, zoom in some more.

As we are accelerating in the car, with constantly changing velocity, we can still have a unique velocity at a particular instant in time.

In mathematical language, for a very small change Δx in either direction from x, we get the same slope, *if* Δx is small enough.

If it's not, we can always make it smaller. That's the beauty of the real numbers.

As you accelerate from 0 to 60, there must be at least one moment in time when your velocity is 50.

Or, put still another way, when they built your house they didn't worry about the curvature of the earth.

If r is the radius of the earth in feet, and the house is $a = 50$ feet long, the drop due to curvature is $r - b$ where $b = \sqrt{r^2 - l^2}$

```
r = 21120000
b = 21119999.9999408
```

That is about 0.00006 feet over the length of a 50 foot house, about a thousand times less than 1/16 of an inch. It is nearly 6 orders of magnitude, one part in a million.

Since the changes in x and y are so small, we use the new nomenclature: dy and dx.

power rule

To actually calculate slopes for curves (and straight lines), use the power rule.

For a horizontal line with zero slope:

$$y = c$$

$$\frac{dy}{dx} = 0$$

For a line with a slope c:

$$y = cx$$

$$\frac{dy}{dx} = c$$

For the parabola, the rule says that if $y = cx^2$, the slope or derivative is

$$\frac{dy}{dx} = 2cx$$

We've been writing c as the constant, so as not to confuse it with a, the acceleration. In analytic geometry, a parabola is usually written with a constant a, called the shape factor:

$$y = ax^2$$

Then, the slope is $2ax$.

If we had

$$y = ax^2 + bx + c$$

with a, b, c all constant, then the slope would be $2ax + b$.

The above uses our three rules from above, plus one more, that when taking the derivative of a polynomial, the derivative of the whole is simply the summed derivatives for each term.

For the equation of motion under gravity

$$s = \frac{1}{2}at^2 + v_0 t + s_0$$

$$v = \frac{ds}{dt} = at + v_0$$

$$\frac{dv}{dt} = a$$

Notice how the 1/2 and the 2 cancel in the second equation.

Continuing to the cubic, if y depends on x^3 like

$$y = cx^3$$

then

$$\frac{dy}{dx} = 3cx^2$$

The general form of the power rule is that if

$$y = x^n$$

then

$$\frac{dy}{dx} = nx^{n-1}$$

The exponent has been reduced by 1 power, and the value of that exponent applied as a factor in front of the expression.

This rule had already been discovered before Newton. It's a toss-up whether Fermat or Cavalieri was first. We will prove this later, but for now we just want to introduce the idea and practice using it.

note

If you already know some calculus you're probably jumping out of your chair while reading this chapter because you've had it pounded into you that dy/dx is not a quotient and believe that you can't simply multiply both sides of the equation by dx.

Well, you can. And I'll explain why as we go along.

Chapter 45

Easy pieces

Integration

Differentiation breaks things up into small pieces dx or dr. Integration adds up many little pieces. The symbol for integration is a relaxed S that stands for summation: $\int$.

As Thompson says

> The word "integral" simply means "the whole." If you think of the duration of time for one hour, you may (if you like) think of it as cut up into 3600 little bits called seconds. The whole of the 3600 little bits added up together make one hour.

We boldly claim that from the point of view of problem-solving, integration is simply the inverse of differentiation.

Mathematicians hate this kind of talk, because it trivializes a profound statement, the fundamental theorem of calculus.

But for practical problem-solving our counter-claim is that this profundity *doesn't matter*. It is also likely to confuse the beginning student, another reason to put it aside for the time being. We'll return to this issue later, when we cover the theory of the subject very lightly.

The sum of a bunch of small pieces dy is equal to the sum of a bunch of small pieces dx times cx, when $dy/dx = cx$ describes how y changes with small changes in x at any particular point.

The key idea is *at any point.* The relationship between dy and dx depends on where you are on the curve. That's why we need integration.

Write

$$dy = f(x)\ dx$$

We want to solve

$$\int dy = \int f(x)\ dx$$

The sum of all the little pieces dy is just y

$$y = \int f(x)\ dx$$

Now, this surely sounds a little vague. But it will turn out that

$$F(x) = \int f(x)\ dx = y$$

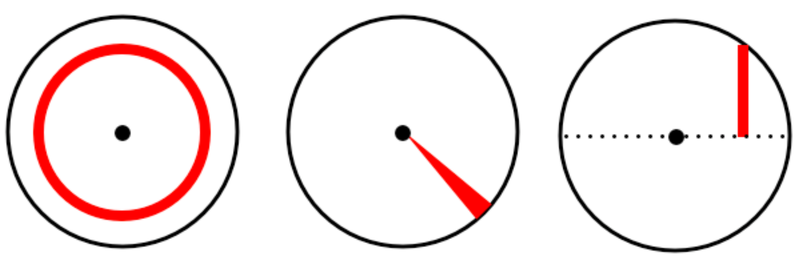
exactly when the derivative of $F(x)$ is $f(x)$:

$$\frac{dF}{dx} = F'(x) = f(x)$$

This is the first of two bright ideas we need to solve an equation like $\int f(x)\ dx$. Just find $F(x)$ such that the derivative of $F(x)$ is $f(x)$.

Area of the circle

Let's spend some time analyzing the area of a circle. This provides crucial insight into what integral calculus can do.



Integration is used to compute areas and volumes, and other sums, by adding up many little pieces.

To calculate the area of a circle, we find the pieces we will use with one of three basic strategies: rings, slices of pie, or rectangles of area underneath the function obtained by solving $x^2 + y^2 = R^2$ (using the positive square root). These three approaches are illustrated in the figure above.

rings

In the first approach (left panel), we imagine the area being computed by adding up the individual areas of a series of very thin, concentric rings.

The total area to be computed is that of a circle of a definite, fixed size, and we denote the radius of this circle by capital R, a constant. On the other hand, the series of rings ranges from the origin of the circle to the circumference of the outmost ring. Each one of this progression of rings has a radius, so we use the lowercase r to describe them, with r being a variable—r varies from 0 at the origin to R at the outside of the circle.

Think about an individual ring, for example the outermost ring, which is similar to the circular peel or rind surrounding a thin slice of lemon. We are working with areas here, in two dimensions, so the slice we imagine to be infinitely thin, and we are working with it as a cross-section or ring.

The area of the ring is the length times the width. The length is the circumference, $2\pi R$ for the outermost ring, but in general, for any of the inner rings it is $2\pi r$. The length is multiplied by the width of the slice, which is a small element of radius, dr. The small element of area contributed by an individual ring is dA:

$$dA = 2\pi r \ dr$$

Another way to explain this equation is to ask the question:

how does area change with increasing radius?

If we take a circle and increase its radius by a little bit, how does the area change? The answer is, it changes in proportion to the circumference, $2\pi r$.

Another way to say the same thing is that the derivative is

$$\frac{dA}{dr} = 2\pi r$$

Proceeding from the first equation, the total area is the sum of the areas for the series of rings.

$$A = \int dA = \int_0^R 2\pi r \ dr$$

It's worth emphasizing how this view is different than the examples of integration one usually sees first in a calculus book: these pieces of area are not rectangles but circles. But it poses most clearly the question we are trying to answer, "how does area change as r changes"?

In order to actually determine a value for the area we need two principles. The first is, as we mentioned before, that the solution to

$$\int f(x) \ dx$$

is $F(x)$ if and only if the derivative of $F(x)$ is equal to $f(x)$.

Continuing with our problem

$$\int 2\pi r \ dr = 2\pi \int r \ dr$$

In this step we used a fundamental rule that a constant can come "out from under" the integral sign. That's not surprising. We already know that (at least in the power rule) the derivative of a constant times some function is that constant times the derivative of the function. We will show that is a general rule later.

Now, we need to find a function whose derivative is r.

$$2\pi \int r \ dr$$

We know that function, it is r^2, with an extra factor of 1/2.

$$= 2\pi \ [\ \frac{1}{2} \ r^2 \ ] \ = \pi r^2$$

Combining all the coefficients we have $\int 2\pi r \ dr = \pi r^2$ precisely because the derivative of πr^2 is just $2\pi r$.

The second principle we need comes from the Fundamental Theorem of Calculus, which takes account of the bounds on the integral (in this case 0 and R). The bounds are written attached to the integral as

$$\int_0^R$$

and on the expression to be evaluated attached to a vertical bar

$$\Big|_{r=0}^{r=R}$$

like this

$$2\pi \int_{r=0}^{r=R} r\ dr = \pi\ r^2\ \Big|_{r=0}^{r=R}$$

We say that the answer is this function, "evaluated between the bounds 0 and R."

The value of such a definite integral is $F(x)$ evaluated at the upper limit minus the value of $F(x)$ evaluated at the lower limit:

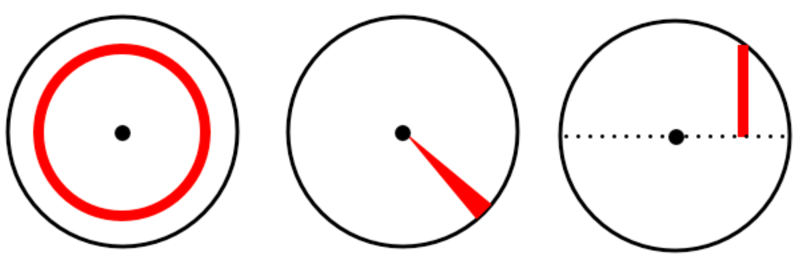
$$= \pi R^2 - \pi(0)^2 = \pi R^2$$

which appears to be correct.

Note in passing that the lower bound doesn't have to be 0, it could be some $\rho < R$. Then we'd have the area of a ring rather than a circle. And another thing, it's not uncommon to leave out the variable from the bounds, and write it like this:

$$2\pi \int_0^R r\ dr$$

wedges



In the second method (middle panel), we need to first find the area of a wedge. For a thin enough slice, this is a triangle, with a familiar formula: one-half the base times the height. The height is R, the radius of the circle.

For the base we need the length of a piece of arc of a circle. Recall that by definition, if we have a unit circle, then the angle of a wedge is equal to the arc it cuts out, and vice-versa, the arc is equal to the angle. (Thus, the total length if we go all the way around the unit circle is 2π).

For a circle with radius R, the length going all the way around is $2\pi R$, and the length of arc for any angle θ is θ times R.

The area we want is built up of a series of wedges that are almost infinitely slender, with angle $d\theta$, so these wedges have bases measuring $R\ d\theta$. The area of each triangular wedge is one-half the height times the base or

$$dA = \frac{1}{2} R\ R\ d\theta$$

For the total area

$$A = \int dA = \int \frac{1}{2} R\ R\ d\theta$$

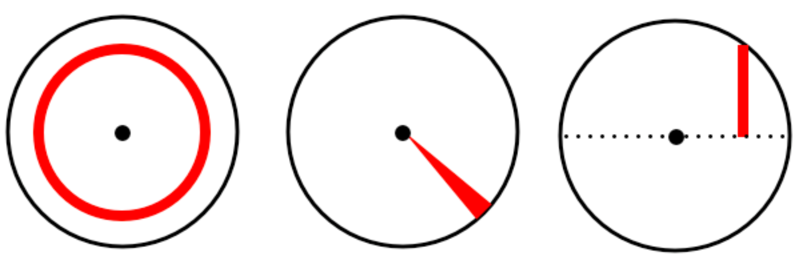
again we see that constants can come outside the integral

$$= \frac{1}{2} R^2 \int_{\theta=0}^{\theta=2\pi}\ d\theta$$

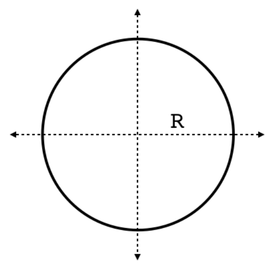
$$= \frac{1}{2} R^2 \theta \ \Big|_{\theta=0}^{\theta=2\pi}$$

$$= \pi R^2$$

area under the curve



The third view (right panel) is the most familiar, but has a somewhat harder calculation. We calculate the area under the positive square root in the equation for a circle (right panel), lying above the x-axis, and then multiply by two to get the whole thing.



$$x^2 + y^2 = R^2$$
$$y = f(x) = \sqrt{R^2 - x^2}$$

To get the area, we need to integrate:

$$\int y \ dx = \int_{-R}^{R} \sqrt{R^2 - x^2} \ dx$$

We will work through this problem **later**, after we review a few more techniques that are useful in doing integration problems.

Of course, the answer will turn out to be just what you'd expect. In fact, this must be so. If we solve the same problem by correctly using two different techniques and get different answers, then at least one of the techniques is wrong.

The area beneath the circle $y = \sqrt{R^2 - x^2}$ and above the x-axis is

$$\frac{1}{2}\pi R^2$$

which is multiplied by 2 to get the area of the whole circle.

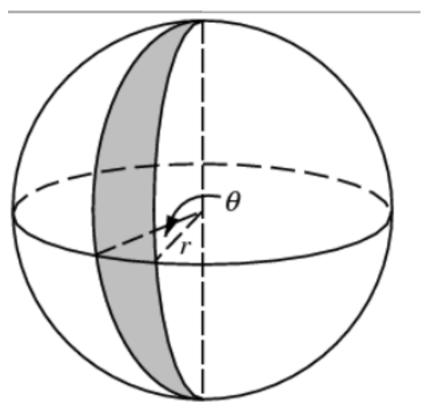
Volume of the sphere

We think about how the volume of the sphere depends on r ($r = 0 \to R$). An incremental change dr changes the volume by adding a thin shell of volume equal to the surface area of the sphere ($4\pi r^2$) times dr. That is

$$
\begin{aligned}
dV &= 4\pi r^2 \ dr \\
V = \int dV &= \int_0^R 4\pi r^2 \ dr \\
&= 4\pi \ \ \frac{1}{3} r^3 \ \Big|_0^R = \frac{4}{3}\pi R^3
\end{aligned}
$$

It's really as simple as that. Of course, you need to know the formula for the surface area to do it that way. Alternatively, if you know the volume of the sphere, taking the derivative is an easy way to get a formula for the surface area.

The image shows a "spherical lune", or segment of the surface of the sphere, as an aid to visualizing the whole surface.



We'll say a lot more about the volume of the sphere **later**.

technical note

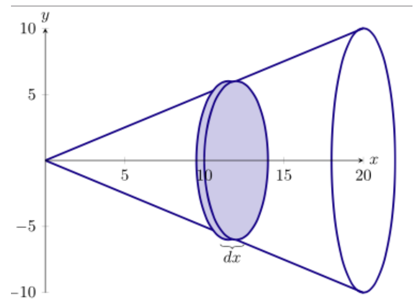
We should point out that this connection between volume and surface area is not true for *every* solid.

As an example, the surface area of a cube of side s is $6s^2$, which would have volume $2s^3$ if the relationship were always correct. In fact, there is something special about the *radial symmetry* of circles and spheres, and their lack of sharp corners and edges.

Here is one more example, to calculate the volume of a cone.

volume of a cone

We lay a cone along the x-axis with its vertex at the origin, opening to the right.



The cone is three-dimensional with the third axis (z) coming up out of the page. The intersection with the xy-plane is a triangle.

Can you see that in the xy-plane y is a linear function of x, i.e. $y = kx$ where k is a constant. The constant k is actually the ratio of the radius R to the height H. That is equal to $\Delta y/\Delta x$.

$$y = \frac{R}{H}x$$

If we slice the cone into thin sections perpendicular to the x-axis, each little piece is a circle with radius y and area πy^2. For a thin enough slice, the volume is that area times the width of the slice:

$$dV = \pi y^2 \ dx$$

Finding the volume of an individual piece is the important part of the calculus argument.

Now we just substitute the value of y in terms of x

$$dV = \pi \ [\frac{R}{H}]^2 x^2 \ dx$$

add up all the little volumes by setting up the integral

$$V = \int dV = \int \pi \ [\frac{R}{H}]^2 x^2 \ dx$$

We apply the basic rule that constant terms can move "out from under" the integral sign:

$$= \pi \ [\frac{R}{H}]^2 \int x^2 \ dx$$

This is a corollary of the result that constants are just carried through in taking the derivative.

We recognize that the value x lies in the interval between 0 and H, $[0, H]$, so these are the "bounds" on the integral, which we write as $\int_0^H$:

$$= \pi\ [\frac{R}{H}]^2 \int_0^H x^2\ dx$$

and then just follow the rule for doing a problem like this: $\int x^2 = x^3/3$. So

$$= \pi\ [\frac{R}{H}]^2\ [\frac{x^3}{3}]\ \Big|_0^H$$

$$= \frac{1}{3}\pi R^2 H$$

This is the answer precisely because the derivative of the result $(x^3/3)$ is equal to the integrand we started with (x^2).

Once again, we obtain the formula of one-third times the area of the base times the height. No matter what the shape of the base is, the area of each slice will be proportional to x^2 and we will end up with a formula involving one-third at the end.

We will see several other methods for obtaining this result.

Note in passing that we can obtain the volume of a frustum (a cone whose top has been cut off) as

$$= \pi\ [\frac{R}{H}]^2\ [\frac{x^3}{3}]\ \Big|_{h_1}^{h_2}$$

$$= \pi\ [\frac{R}{H}]^2\ [\frac{{h_2}^3}{3} - \frac{{h_1}^3}{3}\ ]$$

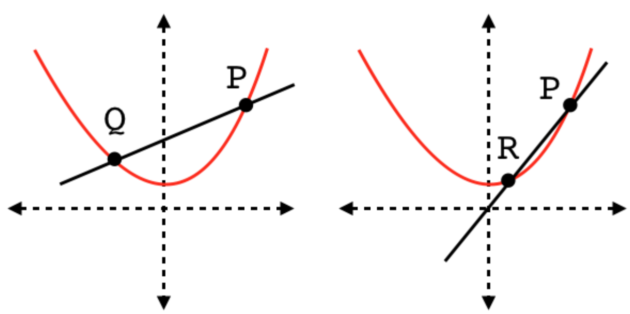
The geometers have given us an even more elegant formula (**here**).

Chapter 46

Difference quotients

In this chapter we look at the geometric interpretation of the derivative — which is the traditional way to begin calculus. The general approach was developed by Fermat.

Think for a minute about a curve such as the one shown in the figure, corresponding to some unspecified function $f(x)$, which looks like it's probably a parabola.



At an arbitrary point P on the curve, for some value of x, we plot $y = f(x)$. This is Descartes' genius idea. The point on the graph of $f(x)$ at x has coordinates $P = (x, f(x))$.

Now consider a point Q near P but also on the curve, For the x-coordinate of Q, a small change is made to x.

We might call that small amount Δx, but many authors use h, a simpler notation, and we will do so as well. The value of the function at $x + h$ is $f(x + h)$ and so Q

has coordinates $Q = (x + h, f(x + h))$.

In this example, h is negative, but that makes no difference. We drew it that way so it's easier to see how the approximation to the slope gets better as we go along.

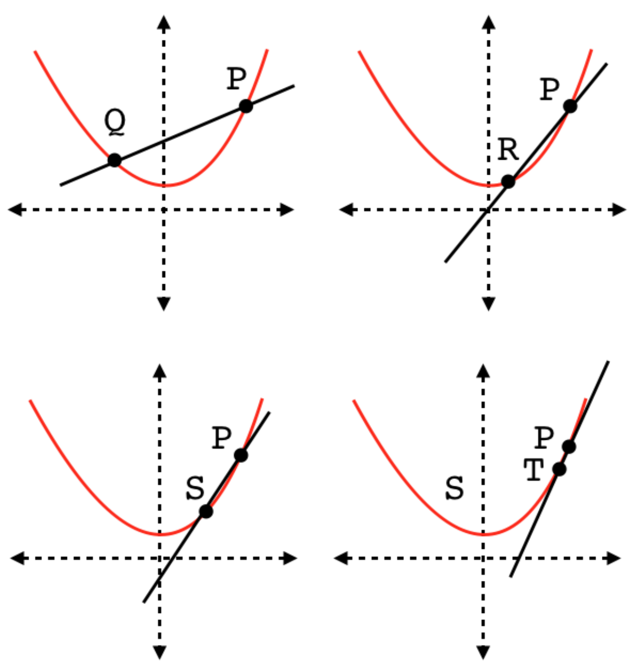
The slope of the (secant) line connecting Q and P is

$$\frac{\Delta y}{\Delta x} = \frac{f(x+h) - f(x)}{x + h - x} = \frac{f(x+h) - f(x)}{h}$$

This is a famous quantity, it's called the **difference quotient**.

The goal of differential of calculus is to find the slope of the *tangent* to the curve at the point P. What we have is an expression for the slope of the secant line PQ, which is close but not quite the same thing.

To go from the secant to the tangent, we ask "what happens to this expression as h gets smaller and smaller and approaches zero." The second point where the secant meets the curve comes closer and closer to the first one.



In mathematical language, we say the slope of the tangent is equal to the limit of the difference quotient as h tends to 0:

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

We'll say a bit more about limits in the next chapter, but for the moment you can think about

$$\lim_{h \to 0}$$

as meaning, "substitute $h = 0$ and see what happens to the expression of interest."

x squared

Let's try a couple of examples and look for a pattern.

$$f(x) = x^2$$

For this function, we write that the difference quotient is

$$\frac{(x+h)^2 - x^2}{h}$$

$$= \frac{x^2 + 2xh + h^2 - x^2}{h}$$

$$= \frac{2xh + h^2}{h}$$

Now divide by the denominator h

$$= 2x + h$$

Finally, to get the slope of the tangent, we evaluate the limit

$$\lim_{h \to 0} \ 2x + h = 2x$$

In evaluating the limit, we ask: what happens to this expression as h approaches 0. In this case, it cannot actually reach zero, because then our previous step of dividing by h would not be allowed. But we let h become really really small, and take advantage of the property of the limit which says that an expression can have a limit at c even if it can't be evaluated at c itself.

At every point on the curve $y = x^2$, the slope of the tangent line to the curve is $2x$. So the slope at $x = 0$ is 0, and the slope at $x = 2$ is 4, and so on.

This process of computing the difference quotient and then finding the limit as $h \to 0$ is called "taking the derivative." It produces an expression which is called the derivative of y with respect to x, in this case

$$\frac{dy}{dx} = 2x$$

and we can interpret this as the slope of the tangent to the curve of $f(x)$ at the point x.

Another useful shorthand uses the f from $f(x)$. We adopt the convention that the derivative of $f(x)$ can be written $f'(x)$.

$$f'(x) = 2x$$

To be even more succinct we might write y' for $f'(x)$.

If we repeat this exercise with a leading constant a (that is, for $f(x) = ax^2$), we find that every term in the numerator of the difference quotient will contain a, and the final result will be $2ax$. Constants just get carried through.

square root

Now look at the square root:

$$f(x) = \sqrt{x}, \quad (x \geq 0)$$

The difference quotient for this function is

$$\frac{\sqrt{x+h} - \sqrt{x}}{h}$$

Clean up the numerator by multiplying by the conjugate

$$\frac{\sqrt{x+h} - \sqrt{x}}{h} \ \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}}$$

$$= \frac{x + h - x}{h\ (\sqrt{x+h} + \sqrt{x})}$$

$$= \frac{h}{h\ (\sqrt{x+h} + \sqrt{x})}$$

$$= \frac{1}{\sqrt{x+h} + \sqrt{x}}$$

We evaluate the limit

$$\frac{dy}{dx} = \lim_{h \to 0} \ \frac{1}{\sqrt{x+h} + \sqrt{x}} = \frac{1}{2\sqrt{x}}$$

inverse

Consider the inverse function

$$f(x) = 1/x, \quad (x \neq 0)$$

$$\frac{\frac{1}{x+h} - \frac{1}{x}}{h}$$

Clean up the numerator

$$\frac{\frac{1}{x+h} - \frac{1}{x}}{h} \quad \frac{(x)(x+h)}{(x)(x+h)}$$

$$= \frac{x - (x+h)}{h\ (x)\ (x+h)}$$

$$= \frac{-h}{h\ (x)\ (x+h)}$$

$$= -\frac{1}{(x)\ (x+h)}$$

We evaluate the limit:

$$\lim_{h \to 0} \ -\frac{1}{(x)\ (x+h)}$$

$$\frac{dy}{dx} = -\frac{1}{x^2}$$

There's a pattern here. We will use the notation $f'(x)$ to indicate the slope of the curve $f(x)$ at x

$$f(x) = x^2 \quad \Rightarrow \quad f'(x) = 2x$$

$$f(x) = \sqrt{x} = x^{1/2} \quad \Rightarrow \quad f'(x) = \frac{1}{2}x^{-1/2}$$

$$f(x) = \frac{1}{x} = x^{-1} \quad \Rightarrow \quad f'(x) = -\frac{1}{x^2} = -x^{-2}$$

The general formula is

$$f(x) = x^n \quad \Rightarrow \quad f'(x) = nx^{n-1}$$

This is easily proved (for integer n) using the binomial expansion for $(x+h)^n$ for integral n ($n \in 1, 2, \dots$). We need only the first three terms:

$$(x+h)^n = x^n + nx^{n-1}h + n\frac{(n-1)}{2}x^{n-2}h^2 + \dots$$

The key point is that the last term shown and all subsequent terms contain powers of h^2 or higher.

After division by h, for each of these terms there will remain one or more terms of h, and in the limit $\lim_{h\to 0}$ these become zero.

$$\lim_{h\to 0} \frac{(x+h)^n - x^n}{h}$$

$$= \lim_{h\to 0} \frac{x^n + nx^{n-1}h + n\frac{(n-1)}{2}\, x^{n-2}h^2 + \dots - x^n}{h}$$

$$= \lim_{h\to 0} \frac{nx^{n-1}h + n\frac{(n-1)}{2}\, x^{n-2}h^2 + \dots}{h}$$

$$= \lim_{h\to 0} nx^{n-1} + n\frac{(n-1)}{2}\, x^{n-2}h + \dots$$

$$= nx^{n-1}$$

Another question is what to do with a sum or difference of polynomials, such as

$$f(x) + g(x)$$

If you write out the difference quotient

$$\frac{f(x+h) - f(x) + g(x+h) - g(x)}{h}$$

everything can be exactly as before, just grouping all terms with $f(x)$ and those with $g(x)$ separately.

$$[f(x) + g(x)]' = f'(x) + g'(x)$$

We showed above by computing the difference quotient directly that

$$f(x) = \sqrt{x}$$

$$f'(x) = \frac{1}{2\sqrt{x}}$$

Here is another approach to the same problem. Consider

$$y = x^2$$

$$\frac{dy}{dx} = 2x$$

Solve for x as a function of y:

$$x = \sqrt{y}$$

We can do algebra with *differentials* (with some constraints):

$$\frac{dy}{dx}\ \frac{dx}{dy} = 1$$

$$2x\ \frac{dx}{dy} = 1$$

$$\frac{dx}{dy} = \frac{1}{2x} = \frac{1}{2\sqrt{y}}$$

In observing the inverse relationship, remember that x and y are related by the equation $y = x^2$. For example, when $x = 2$, $dy/dx = 2x = 4$.

Using the relationship $f(x)$, when $x = 2$, $y = 4$, and $dx/dy = 1/2\sqrt{y} = 1/2\sqrt{4} = 1/4$, which is indeed the inverse of 4.

In this last section, after solving for x as a function of y, y is the *independent* variable. We can switch back to our usual notation:

$$\frac{dy}{dx} = \frac{1}{2\sqrt{x}}$$

problem

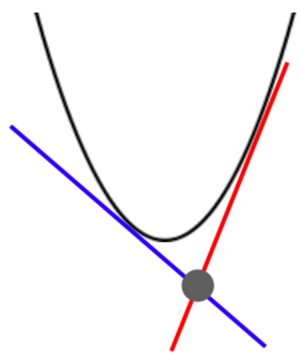
I found the following problems on the web. They are great practice and show what kinds of problems this approach of differentiation can solve. To prove:

Let $(a, f(a))$ and $(b, f(b))$ be two distinct points on the graph of a differentiable function f. Suppose that the tangent lines of f at these two points intersect, and call the point of intersection (c, d). Verifying the following facts is elementary.

1. If $f(x) = x^2$, then $c = (a + b)/2$, the arithmetic mean of a and b.
2. If $f(x) = \sqrt{x}$, then $c = \sqrt{ab}$, the geometric mean of a and b.
3. If $f(x) = 1/x$, then $c = 2ab/(a + b)$, the harmonic mean of a and b.

1

Here is a diagram for the first one:



The claim is that the x-coordinate of the point will be half-way between the x-coordinates for the two points on the parabola. We have:

$$y = f(x) = x^2$$

$$y' = f'(x) = 2x$$

At $x = a$, the slope is $2a$ and the equation of a line through the point (a, a^2) is

$$y - a^2 = 2a(x - a)$$

At $x = b$, the equation is

$$y - b^2 = 2b(x - a)$$

To see where the lines cross, we set the y's to be equal, and solve for x:

$$2a(x-a)+a^2 = 2b(x-b)+b^2$$
$$2ax-a^2 = 2bx-b^2$$
$$2x(a-b) = a^2-b^2$$
$$= (a+b)(a-b)$$
$$x = \frac{1}{2}(a+b)$$

2

We have:

$$y = f(x) = \sqrt{x}$$
$$y' = f'(x) = \frac{1}{2\sqrt{x}}$$

At $x = a$, the slope is $1/2\sqrt{a}$ and the equation of a line through the point $(a, \sqrt{a})$ is

$$y - \sqrt{a} = \frac{1}{2\sqrt{a}}\ (x-a)$$

At $x = b$, the equation is

$$y - \sqrt{b} = \frac{1}{2\sqrt{b}}\ (x-b)$$

We set the y's to be equal

$$\frac{1}{2\sqrt{a}}\ (x-a) + \sqrt{a} = \frac{1}{2\sqrt{b}}\ (x-b) + \sqrt{b}$$

and solve for x. Multiply by $2\sqrt{a}\sqrt{b}$

$$(x-a)\sqrt{b} + 2a\sqrt{b} = (x-b)\sqrt{a} + 2b\sqrt{a}$$

Multiply through and cancel

$$x\sqrt{b} + a\sqrt{b} = x\sqrt{a} + b\sqrt{a}$$
$$x(\sqrt{b}-\sqrt{a}) = b\sqrt{a} - a\sqrt{b}$$
$$= \sqrt{a}\sqrt{b}(\sqrt{b}-\sqrt{a})$$

$$x = \sqrt{ab}$$

3

We have:

$$y = f(x) = \frac{1}{x}$$

$$y' = f'(x) = -\frac{1}{x^2}$$

At $x = a$, the slope is $-1/a^2$ and the equation of a line through the point $(a, 1/a)$ is

$$y - 1/a = -\frac{1}{a^2}\ (x - a)$$

At $x = b$, the equation is

$$y - 1/b = -\frac{1}{b^2}\ (x - b)$$

We set the y's to be equal

$$-\frac{1}{a^2}\ (x - a) + 1/a = -\frac{1}{b^2}\ (x - b) + 1/b$$

and solve for x:

$$(\frac{1}{b^2} - \frac{1}{a^2})x = 2(\frac{1}{b} - \frac{1}{a})$$

$$(\frac{1}{b} + \frac{1}{a})x = 2$$

$$(a + b)x = 2ab$$

$$x = \frac{2ab}{a + b}$$

Part XII

Archimedes

Chapter 47

Value of pi

Archimedes and π

Since Archimedes is a strong presence in this book, we will discuss his method for calculating an approximation to the value of π, the ratio of the circumference of a circle to its diameter.

The commonly cited result is

> The ratio of the circumference of any circle to its diameter is less than 3 1/7 but greater than 3 10/71.

In decimal that is

```
3.140845.. < pi < 3.1428571
```

However, while useful, this misses the main idea: Archimedes described an iterative procedure which can be used to calculate the value of π *to any desired accuracy*.

Although the idea is beautiful, his argument is somewhat unwieldy in detail, so instead we will use modern trigonometry to achieve the same result more economically.

For a discussion of Archimedes actual method (based on a translation by Heath), see this web page

https://itech.fgcu.edu/faculty/clindsey/mhf4404/archimedes/archimedes.html

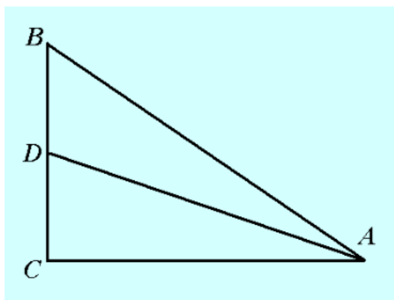
and I have worked out the same proof in detail in the book Best of Calculus.

https://github.com/telliott99/calculus_book

We will also use the trigonometry to find easy formulas for the perimeter and area of inscribed and circumscribed polygons. That part of the argument is partly in this chapter, and the second part follows.

angle bisector

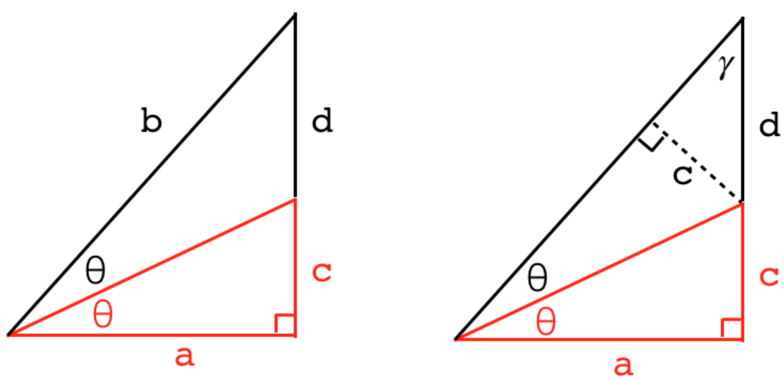
Although we don't follow Archimedes exactly, a key element which he relies upon is the proof that, for an angle bisector in a right triangle, the adjacent sides are in the same proportion as the two segments formed where the bisector meets the other side.



Here:

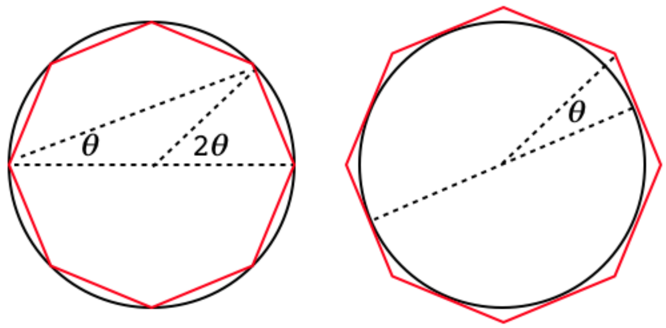
$$\frac{AB}{AC} = \frac{BD}{DC}$$

We showed a proof of this (**here**), or you may be able to reconstruct it from the figure:



the method

We approximate the value of π by squeezing it between the perimeter of an inscribed polygon, which is less than the circumference of the circle, and the perimeter of a circumscribed polygon, which is greater than the circumference of the circle.



The circle of *diameter* has diameter equal to 1 (rather than the radius being 1, which is more usual). The circumference of the circle is then equal to π, the value which gets squeezed between the two perimeters.

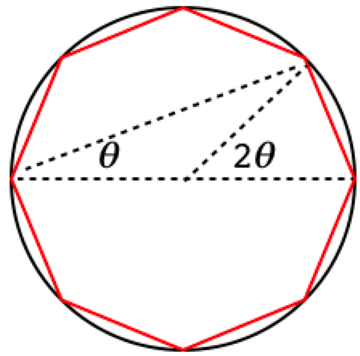
The figure shows a sketch of the polygons when $n = 8$. We will be increasing the number of sides by a factor of 2 at each step, so these are really 2^n-gons with $n = 3$ here.

Finding perimeters in terms of angle θ

For the inscribed circle (left panel), there are 8 sides, so the central angle (marked 2θ) is equal to

$$\frac{2\pi}{8} = \frac{\pi}{4} = 45^\circ$$

and θ is one-half that.

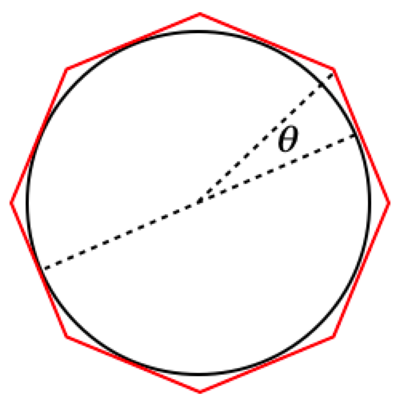


By Thales' theorem, the triangle above containing angle θ, with the diameter as one side, and two other vertices also on the circle, is a right triangle.

The inscribed n-gon side of length S (shown in red) is equal to $\sin\theta$, since the hypotenuse of the triangle is the diameter of the circle, which is equal to 1.

The total perimeter is $8 \cdot S$.

Alternatively, use half the angle at the center of the circle (i.e. θ). Then half the length of the red line $S/2$, divided by the radius ($r = 1/2$) gives $S = \sin\theta$, the same result.



We have the same circle (now showing the outside polygon, circumscribing the circle), it is just rotated slightly. One dashed line extends a bit further to the vertex of the n-gon outside. The angle marked θ is one-half the angle we marked as 2θ previously since now the diameter comes down to the middle of the side.

We compute the whole length of the side T as follows. The half-side is $T/2$ and the hypotenuse of the triangle is one-half the unit diameter, which is $1/2$, so $T = \tan\theta$.

The total perimeter is $8 \cdot T$.

All of this gives us two simple equations for the two perimeters. At each stage there are 2^n sides, the length of each short side S on the inside equals $\sin\theta$ and the length of each short side on the outside T is equal to $\tan\theta$, where $\theta = 2\pi/2^n$.

The total length of the inside perimeter is

$$p = nS = n \sin\theta$$

and that of the outside is

$$P = nT = n \tan\theta$$

When we go from θ to $\theta' = \theta/2$ and $n' = 2n$, we must compute the new values S' and T' from S and T using the half-angle formulas, and then also multiply by 2 to take account of the change from n to $2n$ for the total circumference.

The base case

If we go back to the square ($n = 2$, $2^n = 4$), then the angle θ is $\pi/4$.

The tangent is $T = \tan\pi/4 = 1$ and the sine is $S = \sin\pi/4 = 1/\sqrt{2}$.

Our formulas say that on the inside, the perimeter is $4S = 4/\sqrt{2} = 2\sqrt{2}$ and on the outside, the perimeter is $4T = 4$.

We can just check that from simple geometry. Calculate that the circumscribing square has a side length which is twice the radius of the circle, that is, $s = 1$ for our circle with unit diameter, so its perimeter is 4, which is correct.

Similarly, an inscribed square can be decomposed into four isosceles right triangles with sides of length $1/2$ and hypotenuse $1/\sqrt{2}$, so the total perimeter is $4/\sqrt{2}$, which also checks.

Now, what we will do is to increase n in steps of 1, that increases 2^n by a factor of $2^1 = 2$ each time. Doubling n halves the angle. So all we need is a way to compute trigonometric functions of $\theta/2$, knowing the values for θ, and then we can calculate what happens to the perimeter.

π is simply that value, since the diameter is 1.

Half angle formulas

Quick review (from **here**):

Sine:

$$\sin s + t = \sin s \cos t + \cos s \sin t$$

$$\sin 2t = 2 \sin t \cos t$$

Let unprimed values refer to the whole angle, while primed ones refer to the half-angle. Then:

$$S = 2\ S'C'$$

$$S' = \frac{S}{2C'}$$

Cosine:

$$\cos s + t = \cos s \cos t - \sin s \sin t$$

$$\cos 2t = \cos^2 t - \sin^2 t = 2\cos^2 t - 1$$

$$C + 1 = 2[C']^2$$

Tangent:

$$\frac{1}{T} = \frac{C}{S} = \frac{2[C']^2 - 1}{2S'C'} = \frac{1}{T'} - \frac{1}{S}$$

which can be rearranged

$$T' = \frac{1}{1/S + 1/T} = \frac{S}{1 + C}$$

So, given S, C and T, we can calculate first C' and T', and then finally S'. To get the perimeters, remember that factor of two from doubling n, the number of sides.

Calculation

Let's run a simulation to see what kind of numbers we get. Start with the square ($n = 2$, $2^n = 4$) Previously we found that $S = 1/\sqrt{2}$ and $T = 1$ so

$$p = 2^n S = \frac{4}{\sqrt{2}} = 2.8284$$

$$P = 2^n T = 4$$

Let's try a script to calculate this to larger n.

https://gist.github.com/telliott99/19f521c807210171a4847b319104b3df

Output:

```
> python pi.py
 2 2.8284271247  4.0000000000
 3 3.0614674589  3.3137084990
 4 3.1214451523  3.1825978781
 5 3.1365484905  3.1517249074
 6 3.1403311570  3.1441183852
 7 3.1412772509  3.1422236299
 8 3.1415138011  3.1417503692
 9 3.1415729404  3.1416320807
10 3.1415877253  3.1416025103
11 3.1415914215  3.1415951177
12 3.1415923456  3.1415932696
13 3.1415925766  3.1415928076
14 3.1415926343  3.1415926921
15 3.1415926488  3.1415926632
16 3.1415926524  3.1415926560
17 3.1415926533  3.1415926542
18 3.1415926535  3.1415926537
19 3.1415926536  3.1415926536
>
```

That looks pretty good to me, although it's a bit slow to converge.

This is really quite amazing. Archimedes has not only calculated π to 3 significant figures. More important, he has provided us with an iterative procedure that can be used to calculate the value to *any precision we desire*. As an engineer, Archimedes knew that 3.1416 is precise enough, so he stopped.

After all, no one wants to be William Shanks, or one of these guys:

https://en.wikipedia.org/wiki/Chronology_of_computation_of_pi

Quote:

> [He] calculated pi to [n] digits, but *not all were correct.*

alternative approach to the perimeter

This web page originally got me started with this derivation http://personal.bgsu.edu/~carother/pi/Pi3d.html

(Unfortunately, the link is dead now, probably because the University took Dr. Carother's pages down). It has been preserved by the wayback machine:

https://web.archive.org/web/20171024182015/http://personal.bgsu.edu/~carother/pi/Pi3d.html

On that page, there was given a different pair of formulas, namely, for an inside perimeter p and an outside perimeter P

$$P' = \frac{2pP}{p+P}$$

$$p' = \sqrt{pP'}$$

The first equation can be rearranged to give

$$\frac{1}{P'} = \frac{1}{2}\,[\frac{1}{P} + \frac{1}{p}\,]$$

which is the definition of the harmonic mean of p and P, while the second equation is the geometric mean.

Since in our derivation p and P are the same multiple of S and T, the same relationships should hold for the sine and tangent, but we must remember the extra factor of 2.

From the half-angle formulas, we said that

$$T' = \frac{S}{1+C}$$

Multiply top and bottom on the right by T:

$$T' = \frac{ST}{T+S}$$

Recall that S is the same as p, within a factor of n, and that T is the same as P, within the same factor.

$$p = nS$$

$$P = nT$$

while

$$P' = 2nT'$$

Going back to

$$T' = \frac{ST}{T+S}$$

$$2nT' = \frac{2 \cdot nS \cdot nT}{nT + nS}$$

$$P' = \frac{2pP}{p+P}$$

This is what was given.

□

For the second one

$$S' = \frac{S}{2C'} = \frac{S}{2}\ \frac{T'}{S'}$$

Then

$$2[S']^2 = S \cdot T'$$

$$4[S']^2 = S \cdot 2T'$$

$$[2nS']^2 = nS \cdot 2nT'$$

Changing variables, $p' = 2nS'$

$$[p']^2 = pP'$$

Finally

$$p' = \sqrt{pP'}$$

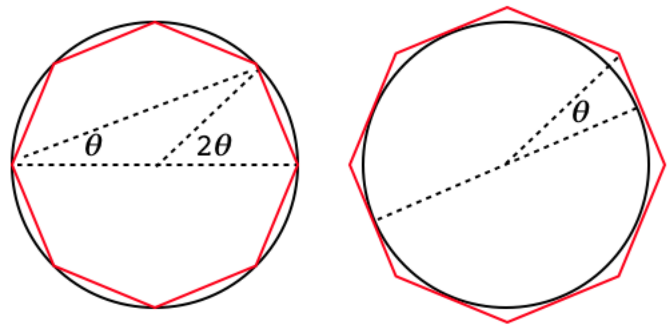
which matches what was given.

□

area

There is yet another way to apply the method, and that is to calculate the *areas* of inscribed and circumscribed polygons. Let's go through this briefly.

For this approach we use a **unit circle** (radius 1) rather than a diameter of 1, as we did above. As before, we define θ to be the central angle of the half-sector (i.e. $\theta = 2\pi/2n$).



Rather than draw an entirely new figure, just imagine in the left panel that we draw the angle bisector of angle 2θ.

Each half triangle has base $\cos\theta$ and height $\sin\theta$, but there are two of them, so the total area of the whole triangle is just $\sin\theta\cos\theta$ and the total area of the inner polygon is

$$a = n\sin\theta\cos\theta = nSC$$

in the notation of this chapter.

As before, to progress to a' we have a factor of 2 as well as the new values S' and C':

$$a' = 2nS'C'$$

For the circumscribed or outer polygon, we just have what we had before, that the side length of the triangle in the right panel is $\tan\theta$ so the total area is

$$A = nT$$

Bring in the half-angle formulas as follows:

$$a' = 2nS'C' = 2n \cdot \frac{S}{2C'} \cdot C' = nS$$

That is slick, but we need an expression for nS:

$$aA = nSC \cdot n\frac{S}{C} = [nS]^2$$

$$aA = [a']^2$$

$$a' = \sqrt{aA}$$

This is like, and yet subtly different than what we had when calculating the perimeter.

Since

$$A = nT$$

and

$$A' = 2nT'$$
$$= 2n\frac{ST}{S+T} = 2\frac{nS \cdot nT}{nS + nT}$$
$$A' = 2\frac{a'A}{a' + A}$$

Compare

$$a' = \sqrt{aA} \qquad A' = 2\frac{a'A}{a' + A}$$
$$p' = \sqrt{pP'} \qquad P' = 2\frac{pP}{p + P}$$

We have primed values in corresponding positions.

However, it turns out that when you take account of the differing size of the circle for perimeter and area methods, and thus the initial values of p, P, a and A, the different order of operations results in precisely the same calculation.

We'll see more in a separate chapter.

Chapter 48

Value of pi revisited

As discussed in a previous **chapter**, Archimedes used paired inscribed and circumscribed polygons to develop an iterative procedure that can be used to calculate the value of π *to any desired accuracy.*

Although the method is beautiful, his argument is unwieldy in detail, so we used modern trigonometry to achieve the same result more economically.

There are, in addition, two other sets of formulas that also reach this end, one based on perimeters, and the other on areas. These formulas are intriguing because they are simple, and it is not surprising that they are connected.

For example, consider a circle of unit *diameter*, so that π is equal to the perimeter. If p and P are the inside and outside perimeters for polygons whose sectors have central angle θ, and the same symbols are used with primes for angle $\theta/2$, then:

$$P' = 2\frac{pP}{p+P}$$

$$p' = \sqrt{pP'}$$

The corresponding formulas for inside (a) and outside (A) areas are (for a circle of unit radius)

$$A' = 2\frac{a'A}{a'+A}$$

$$a' = \sqrt{aA}$$

Notice that these two similar sets of formulas are subtly different. For example, to go from p and P to the primed version, we start with the first formula, while for area we must start with the square root. Part of our purpose in this chapter is to show that this works.

inspiration

It's striking that the formulas for the inside and outside perimeters are so simple, namely $n \sin\theta$ and $n \tan\theta$. The rest just follows from the half-angle formulas.

The web page which originally got me started with the harmonic and geometric mean formulas has been preserved by the wayback machine:

https://web.archive.org/web/20171024182015/http://personal.bgsu.edu/~carother/pi/Pi3d.html

On the very same day that I was revising the previous chapter to better integrate these two approaches, I came across another page which gives a "proof without words" of Gregory's Theorem (that is our subject).

https://divisbyzero.com/2018/09/28/proof-without-word-gregorys-theorem/

It gives these two formulas:

$$I_{2n} = \sqrt{I_n C_n}$$

$$C_{2n} = \frac{2}{1/I_{2n} + 1/C_n}$$

I found this notation a bit awkward, so I substituted the versions given above:

$$a' = \sqrt{aA}$$

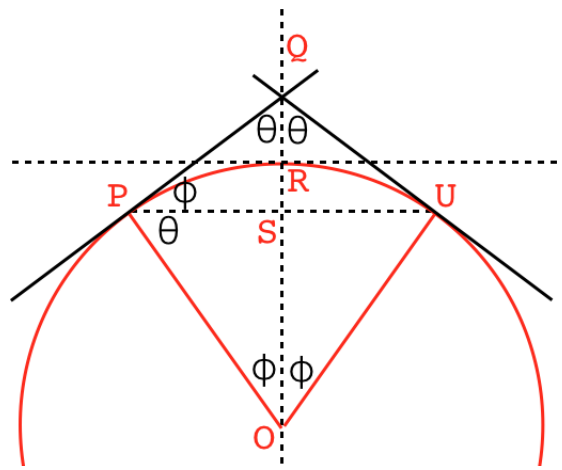
$$A' = 2\frac{a'A}{a' + A}$$

Here, we mainly follow the development from that page and its "proof without words".

One difference is that we will start with the geometry and work backward to the formulas. Another is that we will use quite a few words. Let's deal with the perimeter first and then do the area.

chord of a circle

Before we start the main part, let's just establish some facts about a chord of a circle.



Draw three radii of the circle such that the third bisects the first two, with angles equal to ϕ at the center, point O.

We will show that the chord is parallel to the tangent at R, and perpendicular to OSQ.

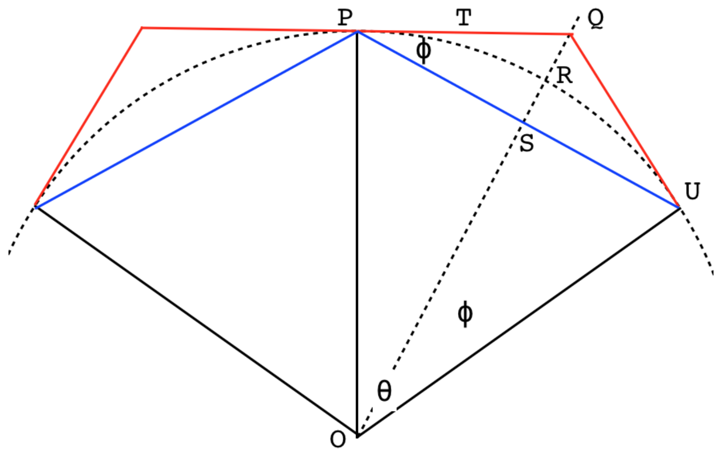
First, tangents are always perpendicular to the radius at the point where they meet the circle.

$\triangle OPS \cong \triangle OSU$ by SAS. Therefore, all four angles at S are right angles. Therefore the chord PU is perpendicular to $OSRQ$. Therefore the chord is parallel to the tangent at R.

The other angle labels are justified as complementary angles in a right triangle.

So any chord which is bisected in this way, is parallel to the tangent above it. And the bisection of the chord follows from the bisection of the base angle at O.

basic setup



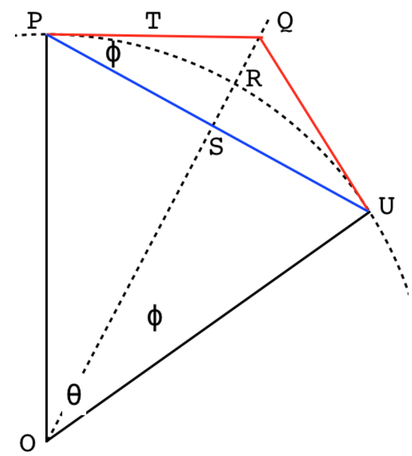
Draw a circle centered at O (only an part of the circle is shown).

Divide the whole 2π radians into $2n$ parts such that ϕ is equal to one of these parts and θ is equal to two of them.

Draw OP, OR and OU as radii, and extend OR to OQ.

Draw the sides of a regular polygon with n sides inscribing the circle and touching its perimeter at P and U. Similarly draw the sides of an n-gon circumscribing the circle and touching its perimeter at the same points P and U.

Two red lines comprise this sector's external perimeter P, while a single blue line is the inscribed perimeter p. The lines of the external perimeter are both tangent to the circle at P and U, and the whole figure is symmetric in each sector, with one blue and two red lines.



By our preliminary results, $\angle PSR$ is a right angle and $\angle SPQ$ is equal to ϕ.

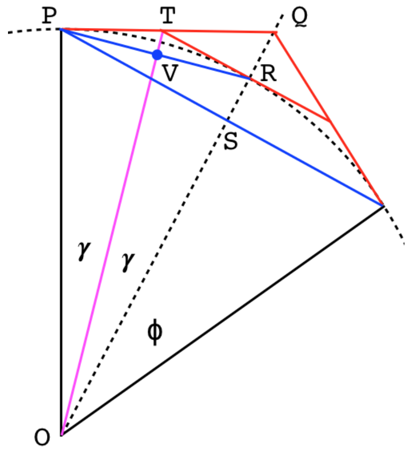
And as we said, the perpendicular bisector to a chord is perpendicular to the tangent at the point where the chord intersects the circle (R).

bisecting ϕ

Next, draw the perimeters p' and P' for the polygon with $2n$ sides and sector angle $\phi = \theta/2$.

It is convenient to rotate the internal perimeter by $\theta/2$ with respect to the external one, a bit to the left when we draw p' and a bit to the right for P'. Both p' and P' touch the circle at R.

A central relationship we use below is that $\triangle PRT$ is isosceles.



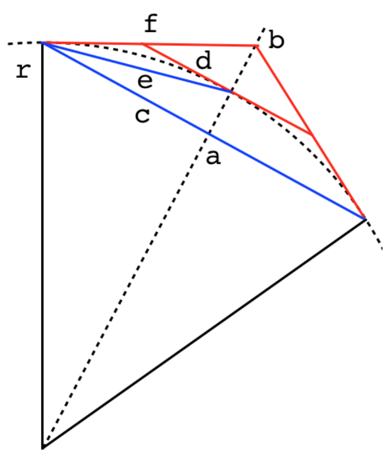
Proof.

$angle VPT \cong \angle VRT$ and the angles at V are right angles, by the preliminary result. Therefore $\angle TPV = \angle TRV$ so $\triangle PRT$ is isosceles. By complementary angles, the base angle has measure γ.

$\square$

It looks as if the segment of the vertical that extends beyond the radius might be equal to that part below down to what looks like the "strut" of a kite. However, this is not true. We will show what this ratio is equal to in just a bit.

Rather than use the vertices as points of reference, let us now label the line segments.



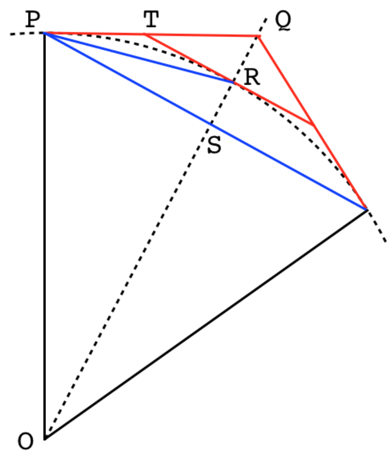
Just to be clear: a is the part of the radius extended to point S in the previous drawing, the intersection of the dashed black and solid blue lines, while b extends all the way to Q, at the vertex of the circumscribing polygon.

c and d are the lengths of the indicated lines *in the half-sector*, not all the way across, and f is the entire length of PQ.

We're ready to proceed.

perimeters

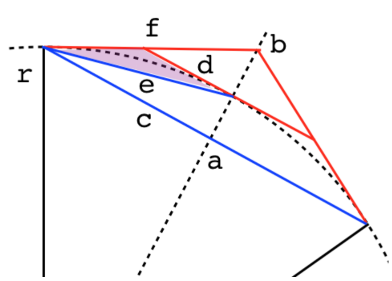
As we said, the key observation is that $\triangle PRT$ is isosceles.



Because of that, and since $\angle SPR = \angle PRT$ by the alternate interior angles theorem, $\angle SPR = \angle TPR$.

Therefore the cosines are also equal, namely:

$$\frac{c}{e} = \frac{e/2}{d}$$



(To see the midpoint of e, drop an altitude in the isosceles triangle, shown in purple).

Therefore:

$$2dc = e^2$$

Now, c is the entirety of p in this half-sector. But d is only one-half of P'.

Hence $2d \cdot c$ is equal to pP', and since $e = p'$, we have that

$$pP' = [p']^2$$

which was our second rule for the perimeters.

The first rule was

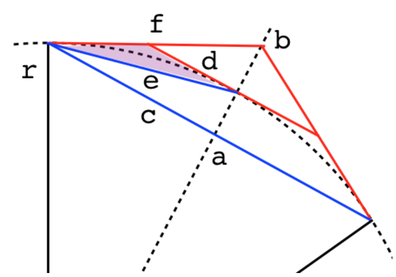
$$P' = 2\frac{pP}{p+P}$$

In geometric terms, we must show that

$$2d = 2\frac{cf}{c+f}$$

$$cd + df = cf$$

Taking another look at the diagram:



The small triangle with base d ($\triangle QRT$ above) has slanted side $f - d$ (subtracting d because, again, $\triangle PRT$ is isosceles). By similar triangles, we have

$$\frac{d}{f-d} = \frac{c}{f}$$

$$df = cf - cd$$

$$cd + df = cf$$

Which is what we needed to prove.

□

areas

The area formulas for inside (a) and outside (A) polygons are those for a circle of unit radius (so that π is the area):

$$A' = 2\frac{a'A}{a' + A}$$

$$a' = \sqrt{aA}$$

This is what we will prove.

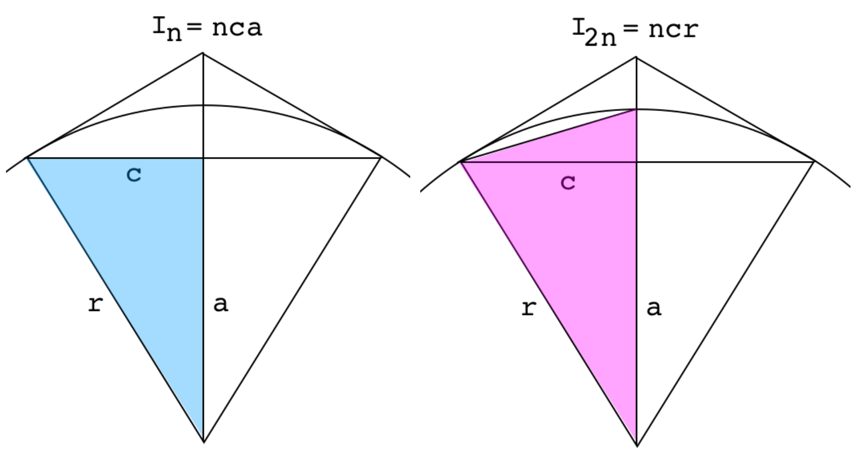
However, having reached this point, we need another symbol for area, because a is currently the line segment corresponding to p/n. Let's use I and C for the inside and outside areas, to match the source.

We will also adopt their n and $2n$ notation, It's a bit clumsy but that will make it easier to match things up. Substituting in the above equations:

$$C_{2n} = 2 \cdot \frac{I_{2n}C_n}{I_{2n} + C_n}$$

$$I_{2n} = \sqrt{I_n C_n}$$

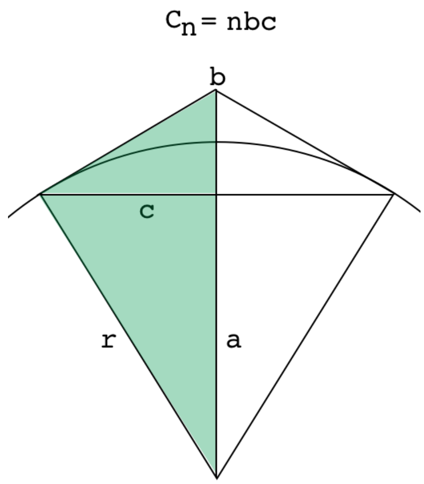
The first two areas are I_n and I_{2n}

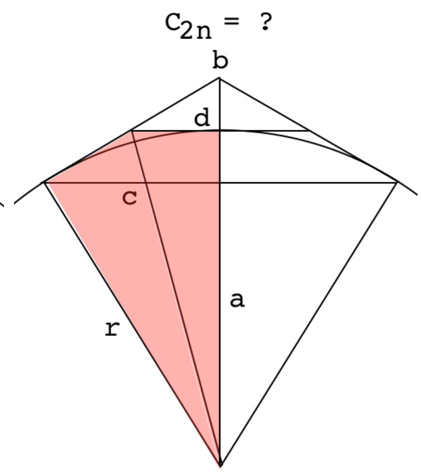


We compute these areas for the whole sector of angle θ, so there are two congruent triangles with base a (or base r) and height c, which makes the factors of one-half go away.

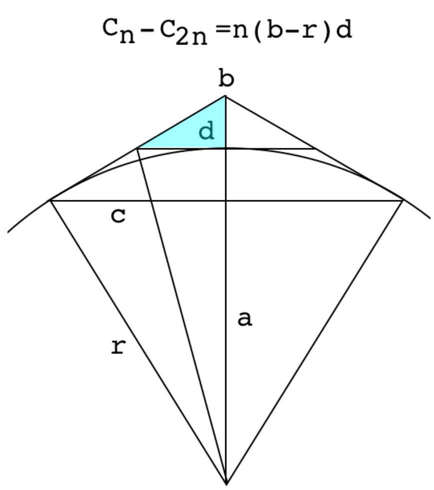
Multiply by n if you like to get the entire polygon, but every expression will have a factor of n, and we'll be looking at ratios, so we can just not worry about it.

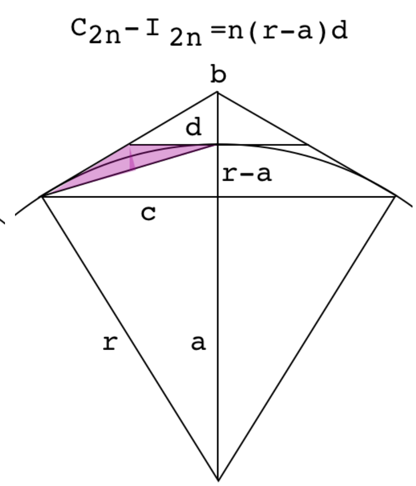
The third easy one is C_n:





We write the last one (C_{2n}) as two different differences.





Let's gather all these expressions in one place, forming ratios:

$$\frac{I_{2n}}{I_n} = \frac{ncr}{nca} = \frac{r}{a}$$

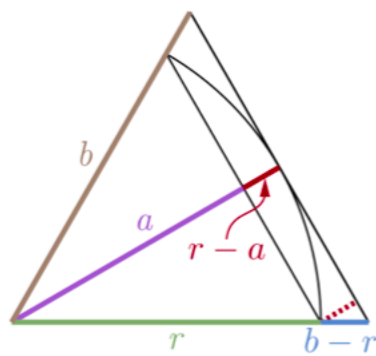
$$\frac{C_n}{I_{2n}} = \frac{ncb}{ncr} = \frac{b}{r}$$

$$\frac{C_n - C_{2n}}{C_{2n} - I_{2n}} = \frac{n(b-r)d}{n(r-a)d} = \frac{b-r}{r-a}$$

We will prove that these three ratios are all equal to each other.

We have used the geometry to prove what the source calls their Lemmas, and those can be used in turn to prove the original Gregory formulas.

But the proof is easy:



It's just a matter of similar triangles:

$$\frac{r}{a} = \frac{b}{r} = \frac{b-r}{r-a}$$

That's the "without words" part.

For that very last part, you can work out the dimensions of the tiny similar triangle, or you can say:

$$\frac{r}{a} = \frac{b}{r}$$

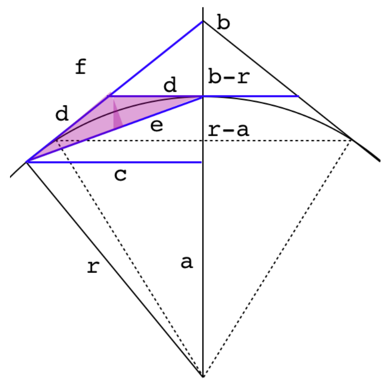
$$\frac{r}{a} - \frac{a}{a} = \frac{b}{r} - \frac{r}{r}$$

$$\frac{r-a}{a} = \frac{b-r}{r}$$

which is easily rearranged to give the desired result.

□

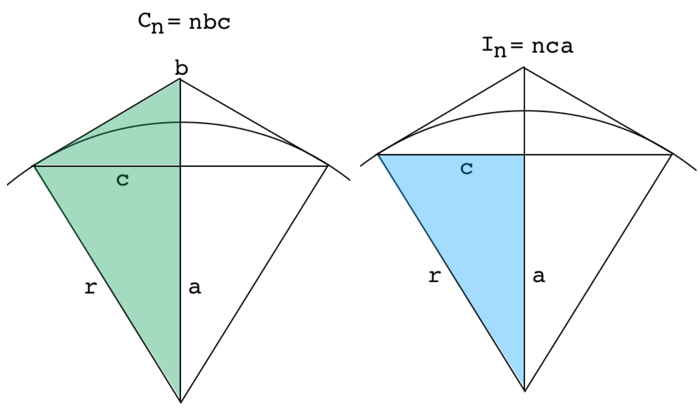
This can also be proved using the **angle bisector theorem**.



The side labeled e bisects the angle formed by the two sides labeled c and f. Therefore

$$\frac{b-r}{f} = \frac{r-a}{c} \quad \Rightarrow \quad \frac{b-r}{r-a} = \frac{f}{c}$$

But f and c are two sides of a triangle which is similar to the colored portions below:



Therefore

$$\frac{b}{r} = \frac{r}{a} = \frac{f}{c} = \frac{b-r}{r-a}$$

As we said.

algebra

Moving on to the geometric mean formula is not hard. From above we have that

$$\frac{I_{2n}}{I_n} = \frac{C_n}{I_{2n}}$$

$$[I_{2n}]^2 = I_n C_n$$

Translated back into the A, a area notation

$$a' = \sqrt{aA}$$

This is just what we wanted to show.

For the other formula, what we have is:

$$\frac{C_n - C_{2n}}{C_{2n} - I_{2n}} = \frac{C_n}{I_{2n}}$$

$$I_{2n}(C_n - C_{2n}) = C_n(C_{2n} - I_{2n})$$

$$2I_{2n}C_n = C_n C_{2n} + I_{2n}C_{2n}$$

$$= C_{2n}(C_n + I_{2n})$$

So

$$C_{2n} = 2 \cdot \frac{I_{2n}C_n}{C_n + I_{2n}}$$

$$C_{2n} = 2 \cdot \frac{1}{1/I_{2n} + 1/C_n}$$

And we're done. In our preferred notation

$$A' = 2 \cdot \frac{1}{1/a' + 1/A}$$

historical note

The area-based formulas given above are due to James Gregory.

https://divisbyzero.com/2018/09/28/proof-without-word-gregorys-theorem/

As an aside, the Fundamental Theorem of Calculus (FTC) is usually thought about (taught and learned) using the language of functions, and ascribed mainly to Leibnitz, with some credit to the two Isaacs, Newton and his university lecturer, Barrow.

https://arxiv.org/abs/1111.6145

Amazingly enough, Gregory published a geometric (Euclidean) proof of the FTC in 1668! That predates Liebnitz (1693) by more than 25 years. This is motivation to give considerable credit to individuals other than Newton and Liebnitz (e.g. Fermat, Pascal, Wallis, Gregory, etc.) in the invention of the calculus.

test

I wrote a simple test of the area formulas using Python.

The script is here:

https://gist.github.com/telliott99/5269b48672cdaeca95c9c9c9d163321d

It gives this output:

```
> python script.py
    4 2.0000000000 4.0000000000
    8 2.8284271247 3.3137084990
   16 3.0614674589 3.1825978781
   32 3.1214451523 3.1517249074
   64 3.1365484905 3.1441183852
  128 3.1403311570 3.1422236299
  256 3.1412772509 3.1417503692
  512 3.1415138011 3.1416320807
 1024 3.1415729404 3.1416025103
 2048 3.1415877253 3.1415951177
 4096 3.1415914215 3.1415932696
 8192 3.1415923456 3.1415928076
16384 3.1415925766 3.1415926921
32768 3.1415926343 3.1415926632
65536 3.1415926488 3.1415926560
>
```

The digits of the output appear to be identical or nearly so. The only difference is that in this script I computed 2^n to give the number of sides. In the previous chapter, we just print n.

details

That's very curious. The first four lines of output from the perimeter version:

```
  2 2.8284271247  4.0000000000
 3 3.0614674589  3.3137084990
 4 3.1214451523  3.1825978781
 5 3.1365484905  3.1517249074
```

and the first five from the area version:

```
 4 2.0000000000 4.0000000000
 8 2.8284271247 3.3137084990
16 3.0614674589 3.1825978781
32 3.1214451523 3.1517249074
64 3.1365484905 3.1441183852
```

It's pretty clear that we are doing the same calculation. It's just that the first column is shifted up by one row.

To confirm that, the perimeter calculation is:

initialization:

$$p = 2\sqrt{2} \qquad P = 4$$

recurrence:

$$P' = \frac{2pP}{p+P} \qquad p' = \sqrt{pP'}$$

The area version is:

initialization:

$$a = 2 \qquad A = 4$$

recurrence:

$$a' = \sqrt{aA} \qquad A' = \frac{2a'A}{a'+A}$$

They give identical results: $A = P$, at each round, but a matches p', or to put it the other way around, p' is retarded by one cycle compared to a'.

Let's try one round of calculation by hand:

$$p = 2\sqrt{2} \qquad P = 4$$

$$P' = \frac{2pP}{p+P} = \frac{2 \cdot 2\sqrt{2} \cdot 4}{2\sqrt{2}+4} = \frac{2 \cdot 2\sqrt{2} \cdot 4}{2\sqrt{2}(1+\sqrt{2})} = \frac{8}{1+\sqrt{2}} = 3.31371$$

$$p' = \sqrt{pP'} = \sqrt{2\sqrt{2} \cdot \frac{8}{1+\sqrt{2}}} = 4\sqrt{\frac{1}{1+1/\sqrt{2}}} = 3.06147$$

The area calculation:

$$a' = \sqrt{aA} = \sqrt{2 \cdot 4} = \sqrt{8} = 2.828427$$

$$A' = \frac{2a'A}{a' + A} = \frac{2 \cdot \sqrt{8} \cdot 4}{\sqrt{8} + 4} = \frac{8}{1 + \sqrt{2}}$$

A' is the same as P'.

The next round for a' is

$$a' = \sqrt{aA} = \sqrt{\sqrt{8} \cdot \frac{8}{1 + \sqrt{2}}} = 4\sqrt{\frac{1}{1 + 1/\sqrt{2}}}$$

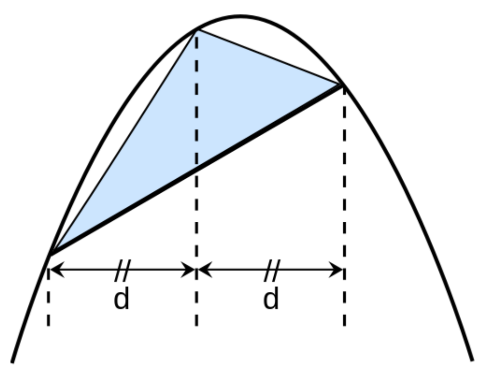
I don't have any words of wisdom to explain why this all dovetails so neatly, but there must be one. When things fit together like this it is never an accident.

Chapter 49

Archimedes and quadrature

Let's talk about Archimedes, and parabolas.

Here is a figure from wikipedia, showing a parabola and a chord of the parabola, which might be drawn between any two points. A triangle is constructed from the chord in the following way: the point dividing the horizontal distance in half is found and that is used for the x-value of the third point.



The Greek genius Archimedes showed that the total area underneath the curve, between the two outside vertices of the triangle, is 4/3 times the area of the triangle shown in blue. The method he used is called the "quadrature of the parabola" and it is (from our modern perspective) a relatively simple though still revolutionary idea.

One very interesting consequence is that the slope of the tangent to the parabola at this midway point is equal to the slope of the chord.

The general equation of a parabola is

$$y = ax^2 + bx + c$$

But for any given parabola, we can translate it to the origin and the parabola at the origin with the same shape is

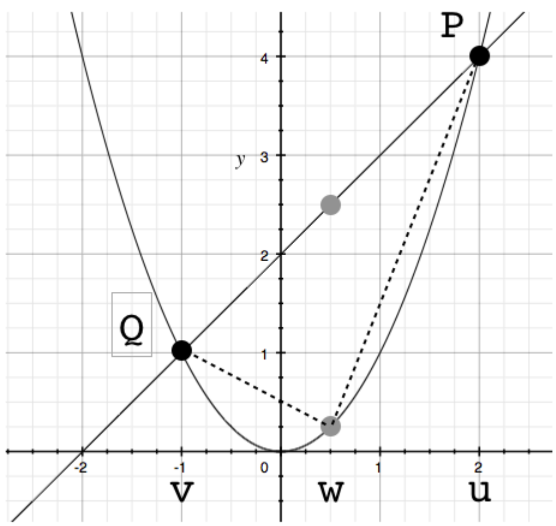
$$y = ax^2$$

This can be demonstrated by completing the square.

If we pick two points on the parabola at $x = u$ and $x = v$, then the corresponding coordinates are

$$P = (u,\ au^2)$$

$$Q = (v,\ av^2)$$



P is the right-hand point in the figure. Let us say that $au^2 > av^2$ and and the slope m of the chord that connects them is

$$m = \frac{au^2 - av^2}{u - v} = \frac{a(u^2 - v^2)}{u - v} = \frac{a(u-v)(u+v)}{u - v}$$

so

$$m = a(u + v)$$

We can see that this formula gives the correct answer for $u = -v$, since the slope at the vertex is 0. Now label the midpoint $x = w$

$$w = \frac{1}{2}(u + v)$$

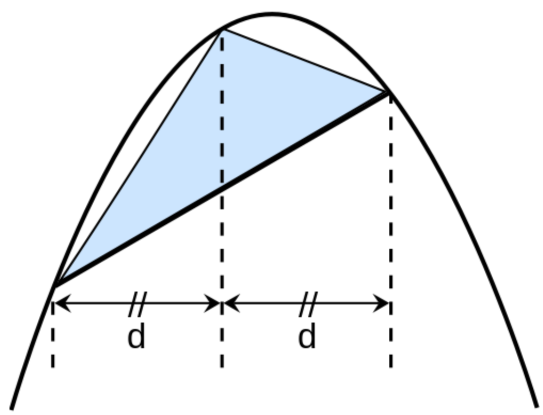
And the slope at w (from calculus) is

$$f'(w) = 2aw == 2a\ \frac{1}{2}(u+v) = a(u+v)$$

So the proposition is correct.

Quadrature

Another interesting thing about this figure is that the area of the triangle can be found from the length of the vertical coming down from the top.



If we simply turn the graph sideways in our mind, then the two small triangles share the part of this line within the blue region, which is their "base", b. And they both have "height" d, since w was chosen as half way between u and v, so their areas are equal, and the total area of the two together is

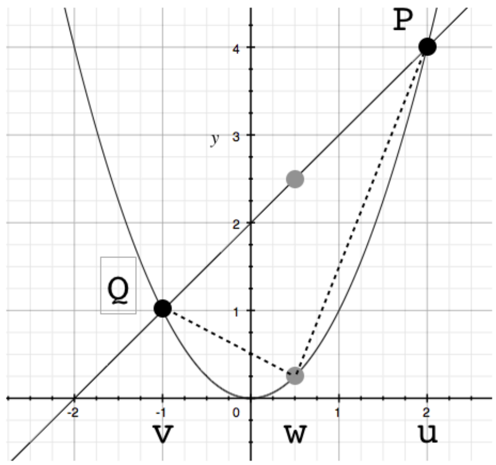
$$A = bd$$

We want to find an expression for the area only in terms of u and v (no b or d). Let's look at the second version of the figure again below.

To repeat, what we found above is that the slope at the point on the parabola corresponding to $x = w$ is equal to the slope of the line that connects v and u, and more important to us now, that the area of the combined triangle (vertices u, v, w) is

$$A = (u-w)\ b = \frac{1}{2}(u-v)\ b$$

where b is the distance parallel to the y-axis between the two points marked in gray.



The length of the "base" b is the average of the y-values for $x = u$ and $x = v$, minus aw^2.

$$b = \frac{1}{2}(au^2 + av^2) - aw^2$$

and from before

$$w = \frac{1}{2}(u + v)$$

so we have

$$b = \frac{1}{2}(au^2 + av^2) - a \ [\ \frac{1}{2}(u + v)\ ]^2$$

Factor out $a/4$

$$= \frac{1}{4}a\ [\ 2u^2 + 2v^2 - (u + v)^2\ ]$$

$$= \frac{1}{4}a\ [\ 2u^2 + 2v^2 - u^2 - 2uv - v^2\ ]$$

$$= \frac{1}{4}a\ [\ u^2 - 2uv + v^2\ ]$$

$$b = \frac{1}{4}a\ (u - v)^2$$

The area is

$$\boxed{A = bd = \frac{1}{8}a\ (u - v)^3} \tag{49.1}$$

check

We'll check three cases to see if this makes sense. First if

$$u = v$$

then the area is zero and $w = u = v$, so that's good. Second, if

$$u = -v$$

then

$$A = \frac{1}{8}a\ (u - v)^3 = \frac{1}{8}a\ (2u)^3 = au^3$$

We compare this result with a direct computation by geometry. In the figure we have two symmetric triangles with individual area

$$\frac{1}{2}u\ au^2$$

The total area is twice that, so it checks. Finally, suppose we have $v = 0$

$$A = \frac{1}{8}a\ (u - v)^3$$

This one is harder to see, but we have that

$$d = \frac{1}{2}(u - v) = \frac{1}{2}u$$

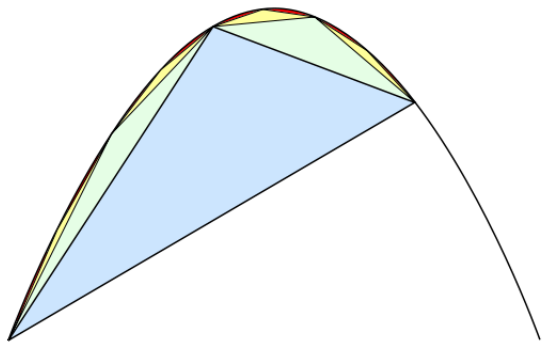
b is the distance between the average y-value which is $(1/2)au^2$ and $aw^2 = a(u/2)^2$

$$b = a\ (\frac{1}{2}u)^2 - \frac{1}{2}\ [\ au^2 - 0\ ]\ =\ \frac{1}{4}a\ u^2$$

$$A = bd = \frac{1}{8}au^3$$

so they all check.

Quadrature of the parabola



The reason for the whole preceding argument is this. The area formula is

$$A = bd = \frac{1}{8}a\ (u-v)^3 = k(u-v)^3, \quad k = const$$

It is solely a function of $u - v$. Suppose we draw two new triangles (in light green, above). For each of these triangles the distance between the new vertices is one-half what we had before. So everything that we have for the big blue triangle is also true for these two new ones, just adjusted by a factor of $u' - v' = (1/2)(u - v)$.

What this means is that the area of each light green triangle is in the ratio to the blue one of $(1/2)^3 = 1/8$. But there are two of these new triangles, so the new area we added is in the ratio $1/4$.

Suppose we do it again, constructing the yellow triangles. The new area of each is in the ratio $(1/4)^3 = 1/64$ but there are now 4 of these yellow triangles so the total area is in the ratio $1/16 = (1/4)^2$

If we call the area of the original triangle T, that of the blue plus the light green is

$$A = T + \frac{1}{4}T$$

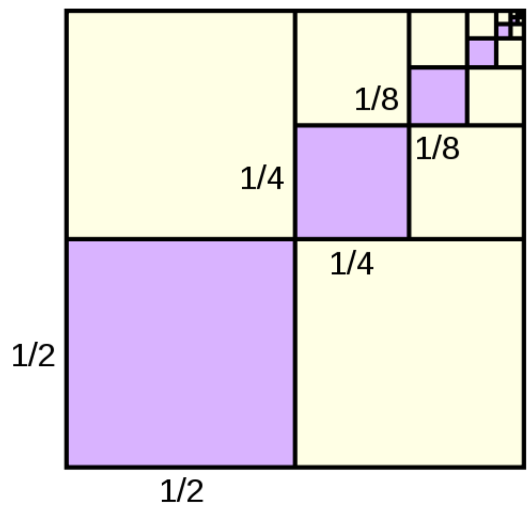
and with the addition of the yellow it is

$$A = T + \frac{1}{4}T + \frac{1}{16}T$$

so, as an infinite series it is

$$A = T(1 + \frac{1}{4} + \frac{1}{16} + \cdots)$$

Here is Archimedes' proof that the sum of this series (not counting the first term) is 1/3.



So the total is 4/3, and the complete area under the parabola is 4/3 the area of the triangle drawn as we described!

This called the "method of exhaustion", and not just because it's a lot of work.

Part XIII

Addendum

Chapter 50

References

- Acheson. *The Calculus Story.*
- Alcock. *Mathematics Rebooted.*
- Corral. *Trigonometry.*
- Courant, Robbins, Stewart. *What is Mathematics?.*
- Densmore. *Euclid's Elements* (annotated Heath translation).
- Dunham. *Euler: The Master of Us All.*
- Dunham. *Journey Through Genius.*
- Dunham. *The Mathematical Universe.*
- Hamming. *Methods of Mathematics Applied to Calculus, Probability, and Statistics.*
- Kiselev. *Geometry, Book I. Planimetry.*
- Kline. *Calculus.*
- Lockhart. *Measurement.*
- Maor. *e, the Story of a Number.*
- Maor. *To Infinity and Beyond.*
- Nahin. *An Imaginary Tale. The Story of* $\sqrt{-1}$.
- Silverman. *A Friendly Introduction to Number Theory.*

- Simmons. *Precalculus mathematics in a nutshell.*
- Spivak. *The Hitchhiker's Guide to Calculus.*
- Stewart. *Significant Figures.*
- Strang. *Calculus.*
- Strogatz. *The Joy of x.*
- Thompson. *Calculus Made Easy.*