The Best of Calculus

Tom Elliott

January 12, 2020

Contents

Part I

Archimedes

Chapter 1

Introduction

This was supposed to be a short book, an exploration of problems like the volume of the cone and sphere, or even just the area of a circle, with some simple physics thrown in. These questions contain within them the heart of calculus: infinities both large and small. I imagine myself looking over Archimedes' shoulder as he explains it to me.

I wrote many of the early chapters 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 — the ideas which make you go ... wow. Now, years later, I still find a lot of pleasure in trying to understand what Kepler and Newton did. It took a genius to figure it out the first time, but it is within anyone's grasp to appreciate what they found.

Then I thought, why not include other favorite problems like the area of the ellipse, the "headlight" problem for the parabola, or the reflective property of the ellipse, and the length and area under the cycloid curve (the "light on a bicycle wheel"). These are problems where calculus easily produces answers that can be checked by more elaborate geometric arguments. In fact, this book might as well be titled .. *Best*

of Calculus and Geometry.

So here we are, with a somewhat longer book.

In the introduction to his book *Calculus*, Morris Kline says

> Anyone who adds to the plethora of introductory calculus texts owes an explanation, if not an apology, to the mathematical community.

I think of this book as akin to ultralight backpacking. We shed weight so as to ascend peaks rapidly, skimming the best of calculus — focusing on geometry and physics, and slinging differentials with abandon. Epsilon is a bit player in the production. Starting with an intuitive notion of adding up many small pieces, we put integrals to work early solving problems.

Going fast allows time to get a view of sophisticated topics, among others, line integrals for work and flux, Newton's proof that a spherical mass acts as a point mass, and integration of a parametrized surface like the torus. Not to mention Kepler's Laws, and a derivation of the Gaussian distribution from first principles.

We do not disdain proof. Proof is central to the enterprise. We prove the Pythagorean Theorem, and the quotient rule for derivatives, as well as Green's Theorem. There is a fun chapter on induction. We prove that π is a constant. In fact, the word "proof" appears nearly 200 times in the text and one of its most interesting features 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 are Morris Kline, Richard Hamming, and Gil Strang. Sylvanus Thompson's simple book is my favorite first text, and it's even a Project Gutenberg project:

https://www.gutenberg.org/files/33283/33283-pdf.pdf

Having said what I like, briefly, here are some things I don't like.

The rigorous approach to calculus pioneered by Cauchy in the 1820's and exported to American schools by Richard Courant in the 1940's is a bad idea. We must motivate rigorous proof by demonstrating utility first. As Ian Stewart says, "proofs come *after* understanding." Courant's method is the way to teach the subject the second or even third time through.

Thompson:

> You don't forbid the use of a watch to every person who does not know how to make one. You don't object to the musician playing on a violin that he has not himself constructed. You don't teach the rules of syntax to children until they have already become fluent in the use of speech. It would be equally absurd to require general rigid demonstrations to be expounded to beginners in the calculus.

A second thing I dislike is calculus problems that are gratuitously arithmetic. Calculus consists of bright ideas, not complicated ones; if the computation is difficult, it's usually *not* a good problem. Also, a good problem often is one with a physical or practical foundation. Having said that, if a course could integrate elementary programming with calculus, I would be very happy.

Finally, a saying attributed to Manaechmus (speaking to Alexander the Great), "there is no royal road to geometry". Which means, practically, that 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 the book on github here:

https://github.com/telliott99/calculus_book

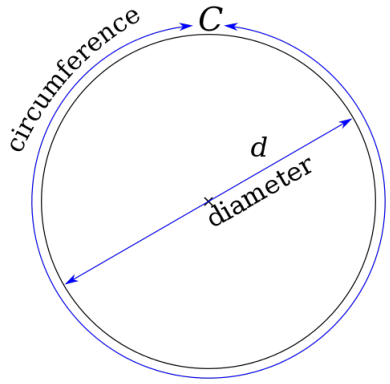
Chapter 2

Area of a circle

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, however, we need to spend some time with circles (also a topic to which he contributed) and look at the problem of the volume of cones and pyramids. These are topics in geometry that come even before the volume of the sphere.

We start with 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.



The proportionality constant is

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

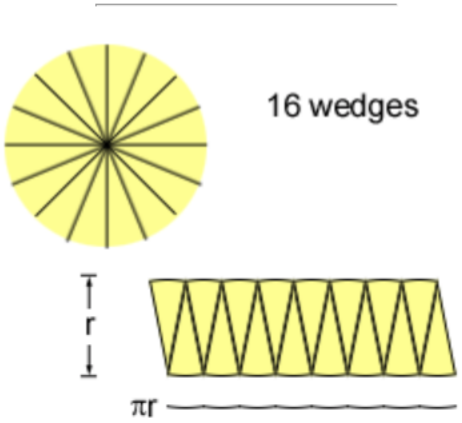
Since the radius is one-half the diameter, $2r = d$ and

$$2\pi r = C$$

This is usually stated as a self-evident fact, but it is actually a theorem to be proved. We will need some of the apparatus of calculus, so we defer the proof.

area of a circle

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, and the length of the long side is approximately one-half the circumference so

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

The trick is to imagine that we subdivide the circle into many slices. If there are infinitely many slices, the edges will be straight and this calculation becomes exact.

According to wikipedia

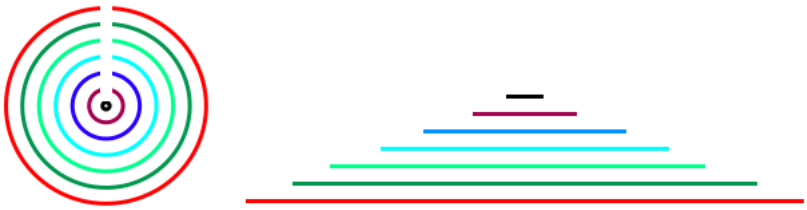
https://en.wikipedia.org/wiki/Area_of_a_circle

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.

Somewhat later, it became clear that for a regular polygon, the area is equal to one-half the perimeter times the altitude from the center to each side (called the apothem). Allowing the polygon to achieve many, many sides, that formula gives $1/2 \cdot 2\pi r \cdot r = \pi r^2$.

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

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



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

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

A 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 this proof, but people have struggled with it, so early in the book. So it will be shown **here**. We retain the quote from Plutarch that follows, it is priceless.

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.

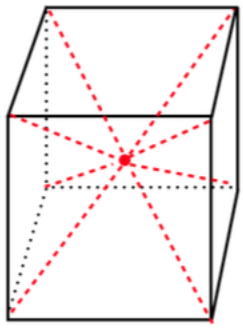
Chapter 3

Volume of a cone

We need a formula for the volume of a cone in order to find the volume of a sphere.

Let's 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.

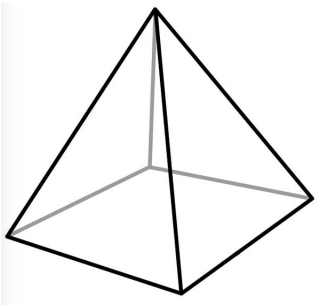


Now we imagine slicing on planes that connect adjacent pairs of lines.

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 is 6 identical pieces (square pyramids) looking something like this



This figure isn't quite accurate because our pyramids will have a height that is $s/2$, 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 the same, or have some of the pieces with equal base and height, but you can't have both.

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

$$6V = s^3$$

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

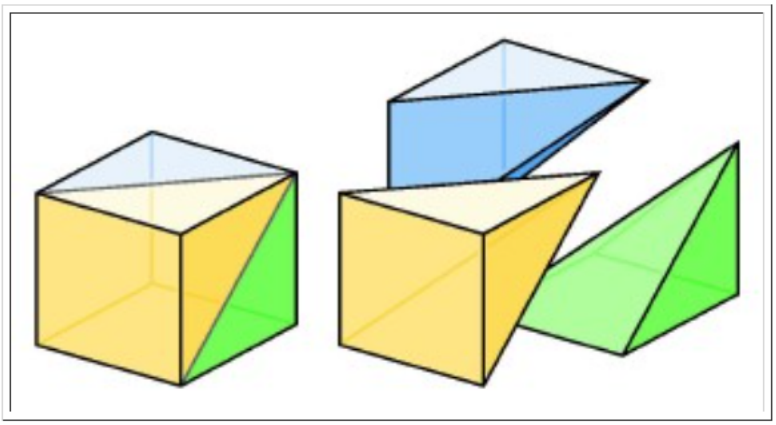
This is the volume for each pyramid with base area s^2 and height $s/2$.

The volume depends linearly on the height and the area of the base. The more general formula for a pyramid is really a linear function of $h = s/2$

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

and you can show this by starting with solids that are longer in one-dimension.

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 3 identical right square pyramids.

The original cube has 12 edges. Each pyramid ends up getting three of those edges, all of them meeting at a vertex, plus it has two more edges along the base where there has been a cut, so the edge was shared.

In addition to those, there are two edges where a cut occurred along the diagonal of a face, and then finally the longest edge is (always the same) interior diagonal of the cube. The total number of edges is 8.

All three pyramids have a single one of the original external (square) bases, two faces that are one-half of an external face cut along the diagonal, and two faces that were originally internal. These latter two faces lie along the plane formed between the original interior diagonal axis and the diagonal cuts of the faces.

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

html

Of course, a pyramid is not a cone. But an argument identical to the one we will use for the sphere shows that the volume is independent of the shape of the base. It just depends on the area. So for a cone we finally obtain

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

I found an algebraic derivation of the factor of one-third, that is given **here**.

We will revisit this problem, to use our first bit of calculus.

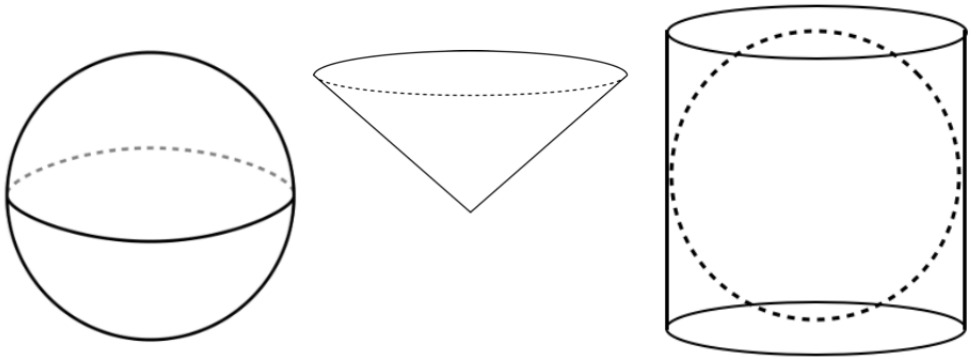
Chapter 4

Archimedes and the sphere

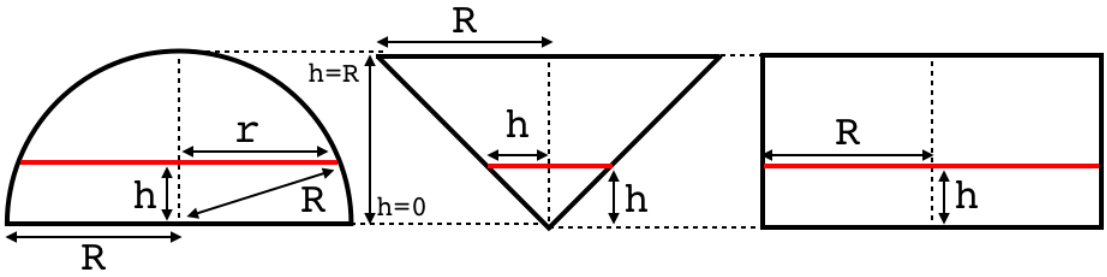
volume of the sphere: geometry

The very first derivation of the volume of a sphere was discovered by Archimedes. The following is his "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 a height h, shown by the red lines. We will choose different values of h later and compare the results, the one shown here is arbitrary.

If you visualize this you should be able to see that each of these red slices is actually a circle. Any cross-section of a sphere is a circle. For the cylinder and cone, cross-sections perpendicular to the axis are circles as well.

The question we ask is: **what is the area for each horizontal 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 areas are proportional to to the radius squared (namely $A =$

πr^2 and so on) and

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

so the areas add too: **sphere plus cone equals cylinder**.

The second 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

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

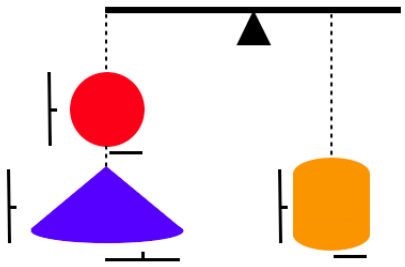
There is a bit of a trick here to hide the idea introduced in calculus, which makes this thinking rigorous. The sphere and cone have variable widths, which means that the radius will be different on the top of a slice compared to the bottom. Therefore, the slices have to be made very thin. In calculus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

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.

I read somewhere that what Archimedes actually balanced is a set-up like that shown here



There are three factors that complicate our calculation: (i) we now have a single cone with radius $2r$ and height $2r$ (because it's doubled in both radius and height the cone's volume is increased by a factor of 2^3), (ii) the sphere and cone are twice as far from the fulcrum as the cylinder, and (iii) the cylinder is made out of something denser than the other objects (four times more dense).

Let πr^3 be one unit of volume, then the volumes are

$$\begin{aligned} \text{sphere} &= \frac{4}{3} \\ \text{cone} &= \frac{1}{3} \times 8 = \frac{8}{3} \\ \text{cylinder} &= 2 \end{aligned}$$

That's $12/3 = 4$ for the sphere plus cone, and furthermore they count double since they are twice the distance from the fulcrum, giving 8 in our volume units. So the left side is $4\times$ the weight on the right side.

However, we are told that the density of the material for the cylinder was four times that of the objects on the left. Hence, it should all balance.

I looked up some densities to try to guess what Archimedes used:

marble	2.56
sand	2.80
copper	8.63
silver	10.40
gold	19.30

How about marble and silver?

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 three sheep or all ten of our fingers. Leopold Kronecker said "God made the integers; all else is man's handiwork".

We will not worry about they come from.

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

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

The brackets contain between them the *elements* of the set. The dots mean that this sequence of numbers continue forever.

It seems hard to know how we can decide whether a particular n is in the set if we can't enumerate all the members of the set. But we

can decide by its form whether 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 at first.

construction of N

To construct the set $\mathbb{N}$, start with the smallest element, 1. Then add successive elements by forming $a_{n+1} = a_n + 1$.

$\mathbb{N}$ is an infinite set, meaning that there is no largest number in $\mathbb{N}$, no largest $n \in \mathbb{N}$.

Suppose $\mathbb{N}$ did have a largest number, a. Well, what about $a + 1$? By the definition we can construct it, and it is clearly also a member of the set, but $a_{n+1} > a_n$ so a_n is not the largest number in the set.

□

set membership

The symbol $\in$ means "in the set" or "is a member of the set".

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 a kind of bound on the numbers. All numbers $n \in \mathbb{N}$ have the property that n is contained in the interval $[1..\infty)$. The right parenthesis means ∞ is *not* part of the interval.

∞ is not a number so it 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 $n \in \mathbb{N}$.

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.

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

$\mathbb{Z}$ stands for the German word *Zahlen*, Number. The set $\mathbb{Z}$ are 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 $p > 0$ with respect to zero using $<$, and all the numbers $n < 0$ using $>$.

inequality

I'm sure you've seen and used the symbols $>$ and $<$, greater than and less than. Among the axioms of the number systems is the collection of *order axioms*. As an example:

- $x < y$ means that $y - x$ is positive
- $y > x$ means that $x < y$

For arbitrary numbers a and b one of three statements is true:

$a < b$, $a = b$ or $a > b$.

There is no 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, which we will talk about soon. Here are just a few more important theorems in this class:

- 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 $ac < bc$

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 usually not $a \times b$, at this level).

And then:

- division a/b, solved by finding a number c so that $c \cdot b = a$.

infinity is not a number

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?

$$\frac{a}{0} = ?$$

As we just said above, this equivalent to finding

$$c \cdot 0 = a$$

But, by definition $c \cdot 0 = 0$.

Suppose we have $c \cdot b = a$ but not $b = 0$. Then, as b gets very small, the number c can get very large. That's OK. We can make c as large as we wish by making b small enough. But we can't say $a/0 =$ something.

If there were such a number (say ∞), then what about

$$\frac{b}{0} = ??$$

$$\frac{c}{0} = ??$$

By definition, we do not allow division by zero.

And we can't answer the question what is $2 \cdot \infty$? If we allowed multiplication by ∞ then the only reasonable answer would be

$$2 \cdot \infty = \infty$$

but then also

$$n \cdot \infty = \infty$$

where n is any number. But then, $2 = n$. This would be a mess.

So, 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 a crucial part of the theoretical basis of calculus.

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

$$\lim_{n\to\infty} \frac{(n+1)-n}{n} = ??$$

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

$$\lim_{n\to\infty} \frac{(1+1/n-1)}{1} = \frac{1}{n} = 0$$

We say that $1/n$ *tends* to zero as n approaches ∞, and so does $[(n+1)-n]/n$.

algebra

As you know, the basic axioms of algebra include the following:

◦ Commutativity for addition and multiplication:

$$a+b=b+a, \quad a\cdot b = b\cdot a$$

- Associativity for addition and multiplication:

$$(a + b) + c = a + (b + c), \quad (a \cdot b) \cdot c = a \cdot (b \cdot c)$$

- Distributivity: $a \cdot (b + c) = a \cdot b + a \cdot c$.
- Additive identity: $0 + a = a$.
- Multiplicative identity: $1 \cdot a = a$.

Chapter 6

Primes

prime numbers

As you know, the positive integers $a > 1$ are of two types. Prime numbers have no factors other than themselves and 1, while composite numbers have at least one other factor. If they are not perfect squares they have two.

The first few primes are:

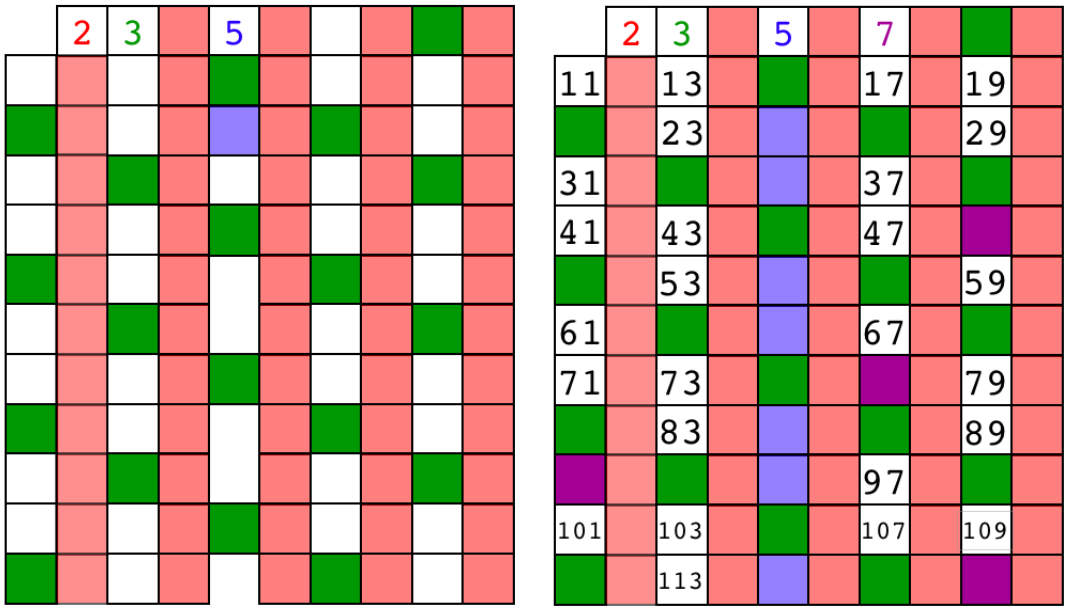
```
2 3 5 7 11 13 17 19 23 29 ...
```

The sieve of Eratosthenes

Eratosthenes is famous in mathematics for his "sieve" which allows one to compute the prime numbers in an economical fashion. We took note of him previously in talking about the circumference of the earth.

The sieve is operated by first enumerating all the integers to some upper limit (here 120). To do things 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.

Starting with the first prime number, 2 (red), eliminate all the numbers divisible by 2 (all the even numbers). Here this has been done by coloring red all of the squares in the even numbered columns (all 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 and 15 go away at this step.

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

The sieve ends when the number for the beginning of the next round, the smallest number not yet eliminated, is greater than the square root of the upper limit (here $\sqrt{120}$). So 7 is used for the last round, 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.

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.

infinite primes

Euclid has a proof that the number of primes (the size of the set of primes) is infinite.

The proof is by contradiction:

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

$$q = (p_1 \times p_2 \times \ldots p_k)$$

$$r = q + 1$$

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

Therefore, none of the known primes divides r, and r is either a prime not in the set of known primes, or the set was originally incomplete.

In either case, the assumption that the set of primes is finite leads to a contradiction.

□

Even for a relatively small number of primes, the second case may hold. Consider

$$(2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13) + 1 = 30031$$

30031 is not prime but is divided by two primes not in the list: 59 and 509.

prime factorization

We will prove that every integer has a unique prime factorization. This is also called *the fundamental theorem of arithmetic*.

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

First, we need a preliminary result, which is called *Euclid's lemma*.

Every natural number $n > 1$, i.e. every positive integer greater than 1, is either prime, or it is the product of two smaller natural numbers a and b.

But the same is true of a and b in turn.

Therefore, every number that can be factored into a and b is the product of the prime factors of a times the prime factors of b.

The notation $m|n$ means m divides n (evenly).

Suppose a given prime p divides $n = ab$, i.e. $p|n$. Then either $p|a$ or $p|b$ (or both).

Proof of existence

The proof is by induction.

Assume the lemma is true for all numbers between 1 and n.

It is certainly true for say, $n \leq 100$, because we can check each case. Start with $n = 101$.

◦ If n is prime (as it is here) there is nothing to prove and we move to $n + 1$.

◦ n is not prime, then there exist integers a and b (with $1 < a \leq b < n$) such that $n = a \times b$.

◦ By the induction hypothesis, since $a < n$ and $b < n$, a has prime factors $p_1 p_2 \ldots$ and b has prime factors $q_1 q_2 \ldots$ so

$$n = ab = p_1 p_2 \ldots q_1 q_2 \ldots$$

This shows that there exists a prime factorization of n.

Proof of uniqueness

To show that the prime factorization is unique suppose that n is the smallest integer for which there exist two different factorizations:

$$n = p_1 p_2 \cdots = q_1 q_2 \ldots$$

Pick the first factor p_1. Since p_1 divides $n = q_1 q_2 \ldots$, by Euclid's lemma, it must divide some particular q_j. Rearrange the q so that q_j is first.

But since p_1 divides q_1 and both are prime, it follows that $p_1 = q_1$.

Now continue the same process with all the factors p_i.

wikipedia:

This can be done for each of the m p_i's, showing that $m \leq n$ and every p_i is some q_j. Applying the same argument with the p and q reversed shows $n \leq m$ (hence $m = n$) and every q_j is a p_i.

□

elegant proof

Hardy and Wright (*Theory of Numbers*, sect. 2:11) have a second proof, which is quite delightful. It is given here almost verbatim:

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

Different factorization:

The same prime P cannot appear in two different factorizations of n, for, if it did, n/P would be abnormal and yet $n/P < n$, the smallest abnormal number.

Thus, we have that

$$= p_1 p_2 \cdots = q_1 q_2 \ldots$$

where the p and q are primes, and no p is a q and no q is a p.

If there exist abnormal numbers with two such factorizations, they must be completely different.

the contradiction

We may take p_1 to be the least p (if the least q is less than the least p, switch labels on all the p's and q's).

Since n is composite, $p_1^2 \leq n$.

The same is true for q_1 and (since $p_1 \neq q_1$), we have that $p_1 q_1 < n$.

Hence, if $N = n - p_1 q_1$, we have $0 < N < n$ and also that N is not abnormal.

Now $p_1 | n$ and since $N = n - p_1 q_1$, so $p_1 | N$.

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

From this 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 $p_2 p_3 \ldots$.

Since q_1 is not a p, this is impossible. Hence there cannot be any abnormal numbers, and this is the fundamental theorem.

□

Chapter 7

Induction

the problem

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

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

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

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

41 is the largest prime smaller than 2000 with this property (I don't know of a proof that no more exist). The known primes that do this 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 $0 \leq n \leq k$, but will never be zero for any other $n > k$. By choosing k large, we generate as many true cases as you like.

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

proving a formula correct

Later in the book, 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 a finite sum like the integers from 1 to n

$$1 + 2 + 3 + 4 + \cdots + n$$

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.:

$$
\begin{array}{ccccccc}
 & & & 1 & & & \\
 & & 1 & & 1 & & \\
 & 1 & & 2 & & 1 & \\
1 & & 3 & & 3 & & 1 \\
\end{array}
$$

Suppose someone has sent us in the mail, a formula for the sum of the first n integers, namely

$$
\begin{aligned}
1 + 2 + \cdots + n &= S_n \\
&= \frac{n(n+1)}{2}
\end{aligned}
$$

Assuming the formula is correct for S_n, then it certainly follows that

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

Rearranging:

$$
\begin{aligned}
&= \frac{n(n+1) + 2(n+1)}{2} \\
&= \frac{(n+1)(n+2)}{2}
\end{aligned}
$$

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

Alternatively, assume the $n-1$ case and prove the formula is correct for n:

$$
\begin{aligned}
S_{n-1} &= \frac{n(n-1)}{2} \\
S_n &= \frac{n(n-1)}{2} + n
\end{aligned}
$$

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

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 then continue all the way to S_n.

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

It must be true for *every* integer n.

This is an example of an inductive proof in mathematics.

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).

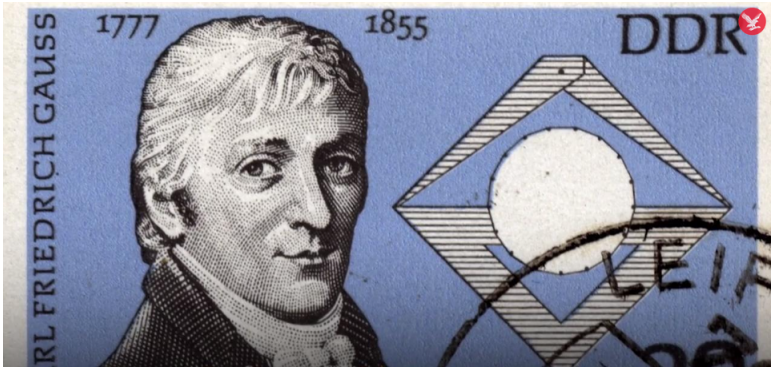
- Graham, Knuth and Patashnik

[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 $n+1$.]

other proofs

Although we proved the formula already by induction, there's no harm in trying a different method.

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.

1	2	...	99	100	S_n
100	99	...	2	1	S_n
101	101		101	101	

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

But in the vertical, we notice that we have n sums, each of which is equal to $n + 1$. So, again

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

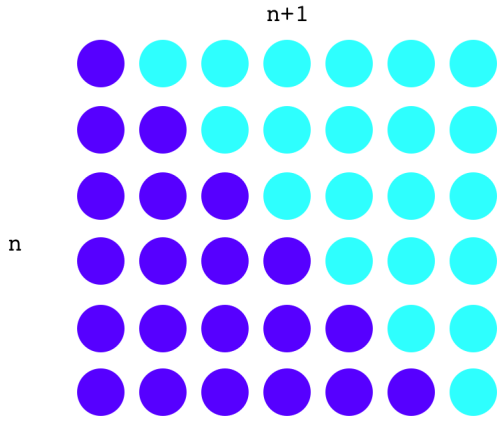
$$S = \frac{1}{2}\, n(n + 1)$$

For $n = 100$ the value of the sum is 5050.

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 $(100+0)/2 = 50$).

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.

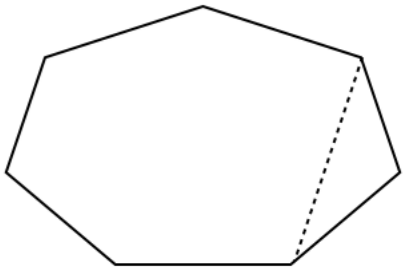


$$2S = n(n + 1)$$

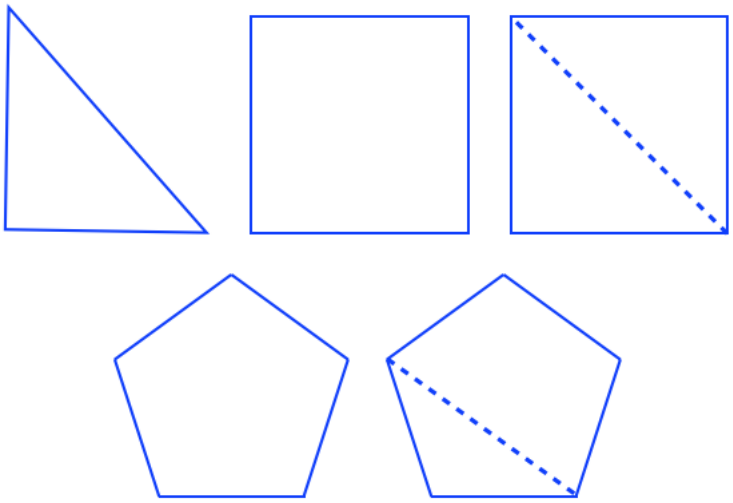
induction in geometry

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 that 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 difference between having n sides and $n + 1$ sides is to add $180°$. 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 $180°$.

$$(n-2)180° + 180° \stackrel{?}{=} ((n+1)-2)180°$$

On the left-hand side, we have the sum of angles for n sides, which we assume is correct, and then we just add $180°$ to it. On the right, we have substituted $n + 1$ into the formula.

Now we need to show that these are equivalent. But of course

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

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

□

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:

$$234783738 \Rightarrow 2+3+4+7+8+3+7+3+8$$

$$= 5+11+11+10+8$$

$$= 16+21+8$$

$$= 7+3+8 = 9$$

Yes, $9|234783738$.

We propose that

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

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.

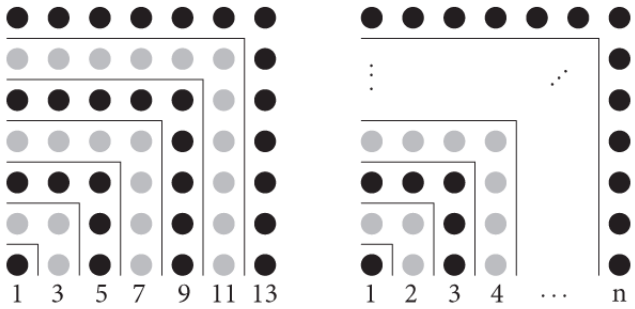
□

Given this, it is easy to show that the sum of digits method always works. I'll leave it as an exercise.

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.

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

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

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 $2 \times (n + 1) - 1$ 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

$$1 = 1^2$$

□

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.

□

Part III

Lines and triangles

Chapter 8

Euclid

Euclid and the postulates

Greek geometry starts hundreds of years before Euclid, whose life overlapped (on both ends) that of Alexander the Great. We know that Euclid died about 270 BC, although today his life and death are shrouded in mystery.

Euclid's book *Elements* is still an excellent place to begin surveying the foundations of geometry. It is really a textbook, a collection of everything that was known about the subject at the time.

Euclid's geometry is mainly constructions (geometric figures) drawn with a pencil on a flat piece of paper, using a straight-edge or a compass or both.

Here are Euclid's first three *postulates* — statements that are 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 radius and one endpoint as center.

Let us assume these as well.

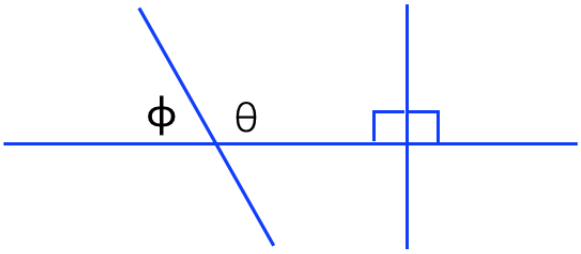
We finesse the difficulty in defining what is meant by *straight* in the real world. If you've ever done any carpentry, for example, you probably know that unknown edges are determined to be straight by comparison with known straight edges. But the mental construct of "a straight line is the shortest distance between two points" gets around this limitation.

The fourth postulate is:

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

This one prompts a different question: what is a "right angle"?

If one line is drawn crossing a second one, let us refer to the two angles formed on one side of the line as supplementary angles (sometimes called adjacent angles)



On the left, one of the angles, ϕ, is smaller than the other one, θ. Alternatively, on the right, the two angles have equal measures. The third possibility, the mirror image of the first, would have $\theta < \phi$.

The definition of a right angle is that

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

An immediate consequence is that the total of all the angles around a point is equal to four right angles.

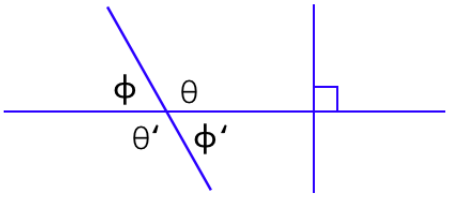
A right angle is frequently designated by drawing a small square at the intersection. Since both angles are right angles, only one square is needed or usually drawn.

In all cases, the sum of the two angles $\phi + \theta$ is equal to two right angles or 180 degrees. There is nothing particularly special about 180 degrees for two 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 one whole turn is equal to 1.

Later, we'll see 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.

Now, consider those angles below the horizontal



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

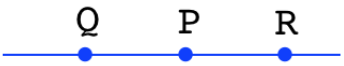
$$\phi + \theta = \theta + \phi'$$

We conclude that $\phi = \phi'$ and $\theta = \theta'$. On the right, if any one of the angles where two lines cross is a right angle, then all four are right angles.

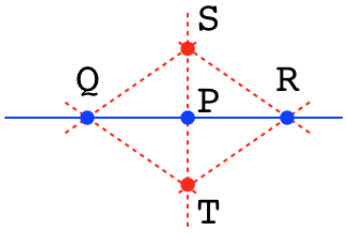
This is called the *vertical angle theorem.*

erecting a perpendicular

There is a simple method to construct a line segment perpendicular to a line at a particular (given) point, or alternatively, through any point not on the line



Consider the blue horizontal line and suppose we know a point on the line P and wish to construct the vertical line through P. The procedure is to use the compass to mark off points Q and R on the line an equal distance from P.



Using the compass again, find S and then draw the line segments such that $QS = RS$. The line segment SP will be perpendicular to line containing QPR.

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 ($QS = RS$). Also find T such that $QT = RT$. The line segment ST will be perpendicular to line containing QR, and pass through S, as required.

By the end of the chapter you will be able to prove this works.

Also, see the video at the url:

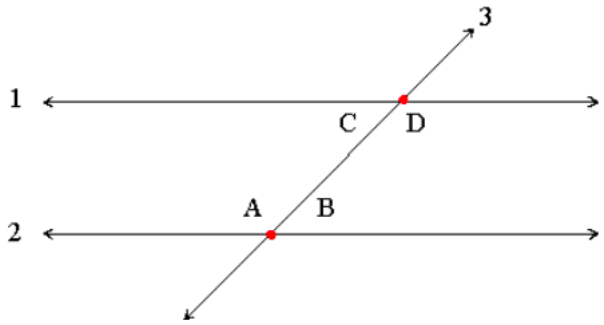
https://www.mathopenref.com/constperpextpoint.html

parallel postulate

This seems rather obvious.

The fifth and final postulate is more subtle:

∘ 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.



Line 1 and line 2 are parallel, if and only if, $A+C = B+D = 180 = 2$ right angles. This postulate is equivalent to what is known as the parallel postulate.

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

But we also know from the properties of two lines given above that $A + B = 180$ degrees. So

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

and then

$$C = B$$

This is called the theorem on *alternate interior angles*.

Consider a familiar situation where this is not true. Suppose we are doing geometry on the surface of a sphere, such as the earth. 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 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 that seem unobjectionable. 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.

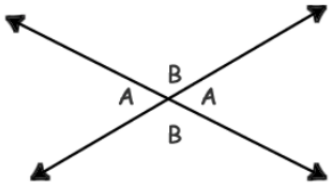
We will see how to proceed from the postulates and axioms to various proofs. *Given these assumptions*, we can prove theorems that must be true.

Thales

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

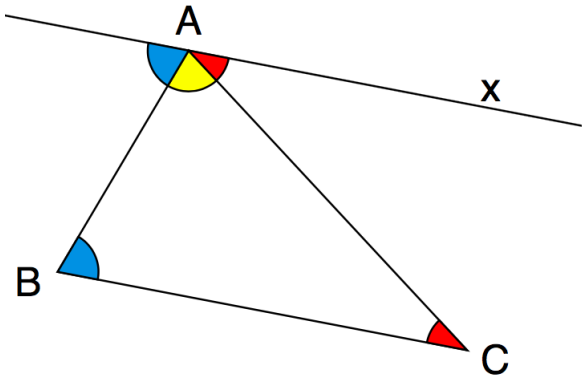
Dunham has written a lot about the history of mathematics in Greece, starting with Thales (624-546 BC), who was from a Greek town called Miletus on the coast of Asia Minor (modern Turkey). He lived long before Euclid. Although none of his writing survives, it is believed that Thales proved several early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

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



This theorem 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."

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

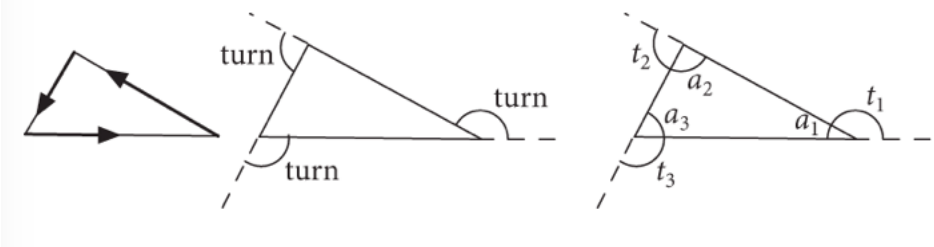


This theorem depends on the theorems we developed above. Draw a line segment through A parallel to BC (by doing a vertical construction twice. Now, use alternate interior angles and follow the colors to the result.

another proof

Here is a different proof of the theorem on the sum of angles in a triangle adding to 180 degrees..

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, t, called the exterior angle. After passing through all three vertices, we must end up facing in the same direction as we started.



$$t_1 + t_2 + t_3 = 360$$

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

$$(t_1 + a_1) + (t_2 + a_2) + (t_3 + a_3) = 3 \cdot 180 = 540$$

By subtraction

$$a_1 + a_2 + a_3 = 180$$

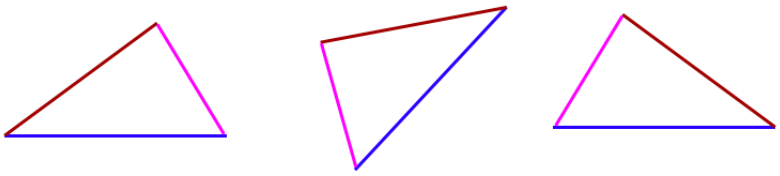
Chapter 9

Congruent triangles

Congruence and similarity

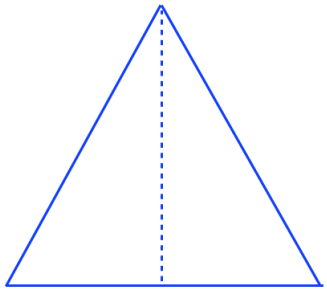
∘ Two triangles are *congruent* if and only if they have the same three side lengths. This is often abbreviated SSS (side-side-side).

By this definition, a triangle and its mirror image are congruent. The three triangles shown below are all congruent, even though the first is flipped (it is the mirror image of the other two).



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

In this figure the two smaller triangles obtained by dividing an equilateral triangle in half, are congruent.



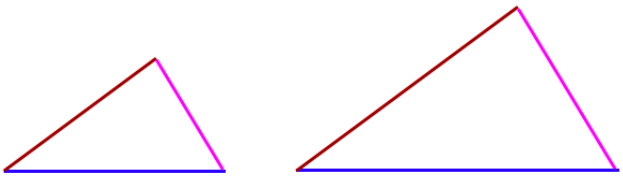
We will prove this at the end of this chapter.

Some triangles are *similar* but not congruent, with all three angles the same but of different overall sizes. We could call this AAA (angle-angle-angle). For similar triangles, the three corresponding pairs of sides are in the same proportions, but re-scaled by the same constant of proportion.

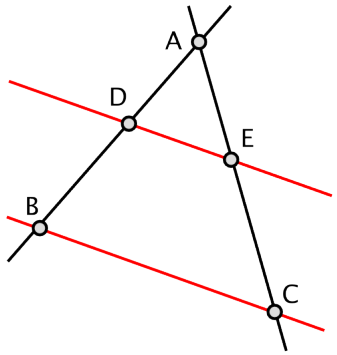
∘ Two triangles are similar if they have the same three angles.

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.

Similar triangles have their sides in the same proportions.



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. Similarity means that all the angles are equal. This is easily proved using the theorem on alternate interior angles.



In this example, these ratios are all equal

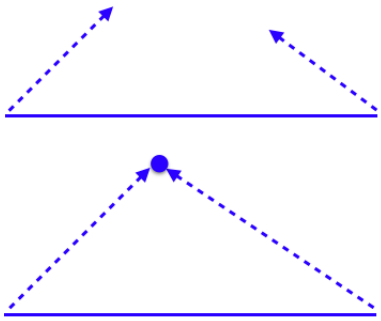
$$\frac{AD}{AB} = \frac{AE}{AC} = \frac{DE}{BC}$$

In addition to SSS (side-side-side), there are 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)

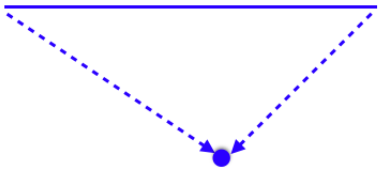
constructions

Again, the way I think about these 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:



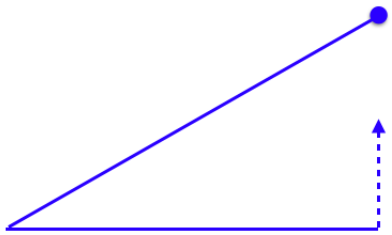
Plot the known side and start two other sides from the ends of that side containing the known angles. They must cross at a unique point.

But... actually, if we start the two lines pointing below the horizontal, there is another solution, the mirror image. This triangle is also congruent to the one above.



Alternatively, knowing two angles means 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 (this part is important) we also know *which* two angles flank the known side.

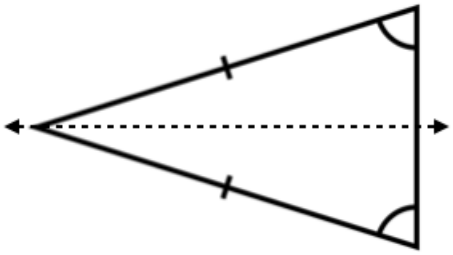
For a right-triangle, if the hypotenuse and one leg are equal, the two triangles are congruent.



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.

another theorem from Thales

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

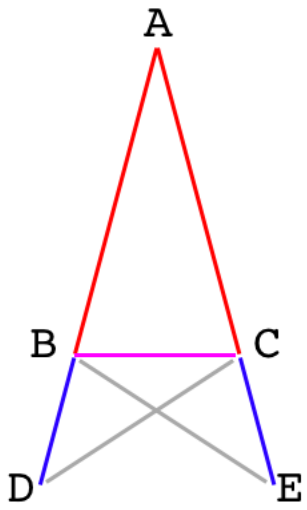


My favorite proof of this theorem is from symmetry (above). Draw a line from the vertex between the two equal sides to the midpoint of the base opposite. If you turn the triangle over along this axis, we obtain the same triangle back again.

Alternatively, just say AAS or use the previous theorem on right tri-

angles.

Euclid's proof is here:



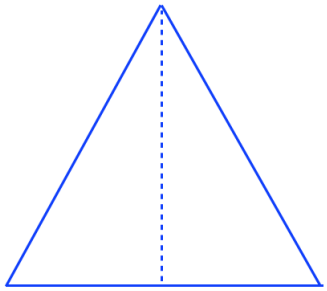
AB = AC (given)
AD = AE (given)
ΔABE = ΔACD (SAS)

BD = CE (subtraction)
BE = CD (congruent Δ)
ΔBCE = ΔBCD (SSS)

∠BCE = ∠CBD (congruent Δ)
∠ABC = ∠ACB (subtraction)

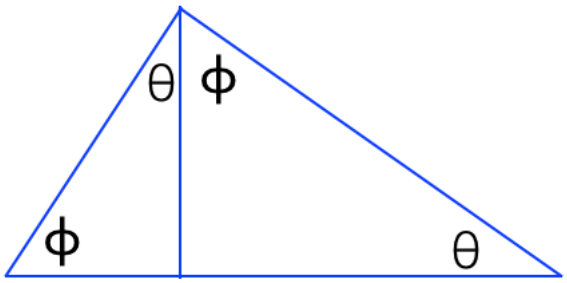
The theorem says that the base angles are equal $\iff$ the two sides sides (not the base) are equal. We proved that equal sides lead to equal angles, now we must proceed backwards, from equal angles to equal sides. Sometimes proofs can get a little complicated, so I've put that in a separate chapter **here**.

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.



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 the two angles at the top where the altitude meets the sides are also equal (as the third angle when the other two angles are known).

Therefore the two smaller triangles are congruent, by side angle side (S-A-S)



In any right triangle, if an altitude is dropped to the long side (called the *hypotenuse*), then the two smaller triangles that are formed are similar to the original one.

The reason is that in any right triangle the two angles that are not the right angle add up to 90 degrees, so the whole will be 180 degrees. They are called *complementary* angles.

Each smaller triangle contains an angle also found in the large one. Therefore the opposing, complementary angles are equal.

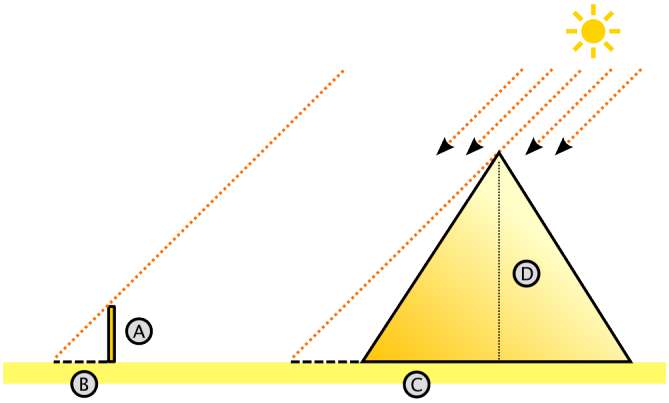
pyramid height

As we said, Thales was from Miletus and he lived around 600 BC. Thales is believed to have traveled extensively around the Mediterranean 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 the Romans for colonies, and their major city, Carthage, was destroyed much later by the Romans in the third Punic War.

During his travels, he 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.



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.

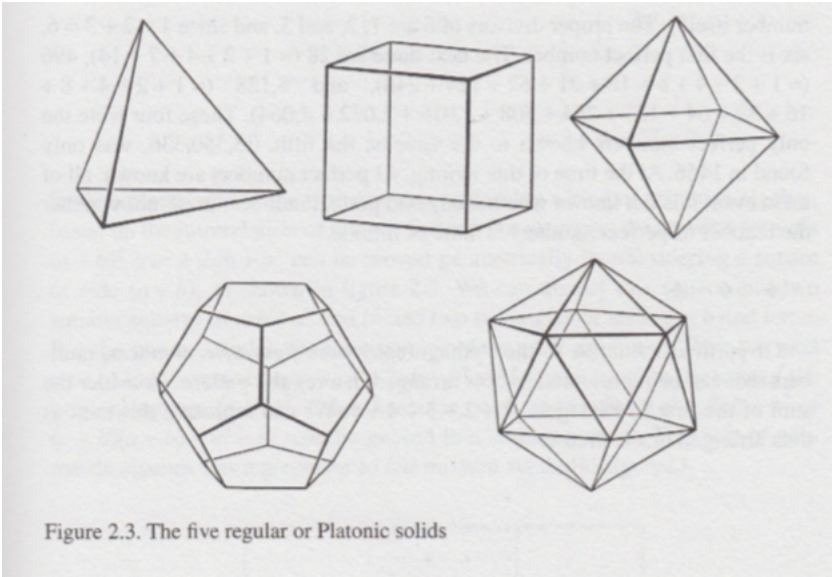


Figure 2.3. The five regular or Platonic solids

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

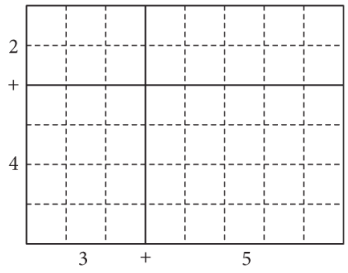
Chapter 10

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? Well, every triangle can be turned into a parallelogram, by attaching a rotated image of itself, like this:

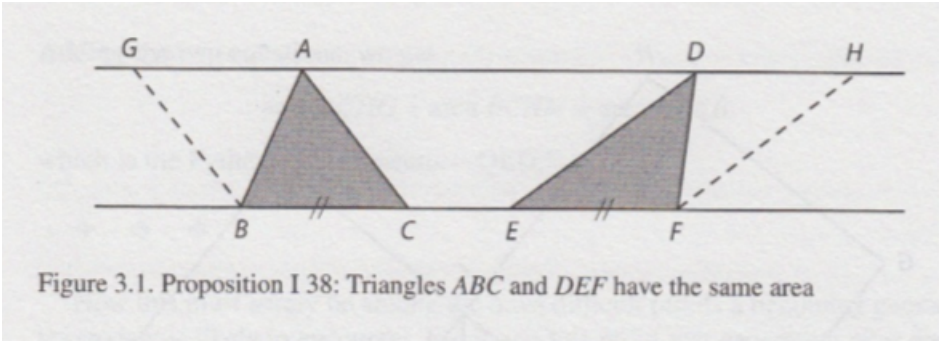
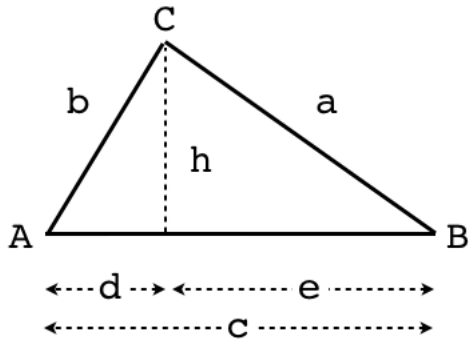


Figure 3.1. Proposition I 38: Triangles *ABC* and *DEF* have the same area

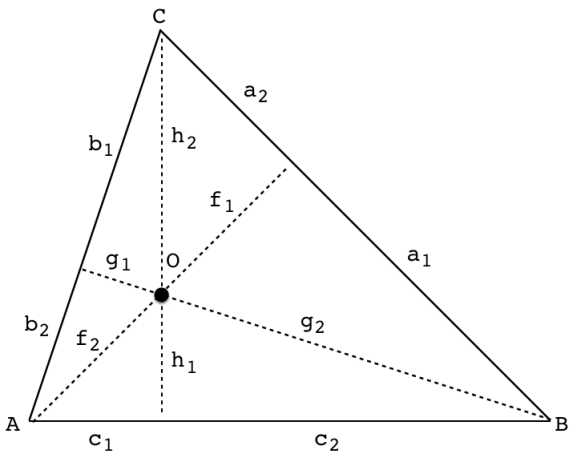
An acute triangle is on the left and an obtuse triangle on the right. Since the area of each triangle is one-half that of its corresponding parallelogram (because we added the same area to make the parallelogram), the area of a triangle is one-half the base times the height.



Here, the area is $hc/2$.

We could choose any side of the triangle to be the base and then multiply $1/2 \times$ base $\times$ 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:



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

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

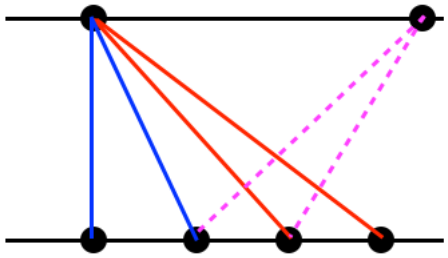
But each of these smaller areas can be computed in different ways. In particular $\triangle BOC$ can be viewed as having base g_2 and height b_1, while $\triangle AOB$ can be viewed as having base b_2 and height g_2, so (twice) the total area is also

$$\begin{aligned} 2A &= b_1 g_2 + b_2 g_2 + bg_1 \\ &= bg_2 + bg_1 = bg \end{aligned}$$

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.



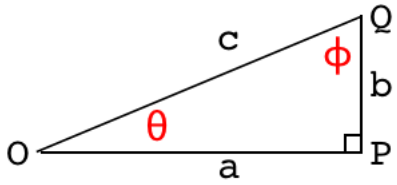
Chapter 11

Angle bisector

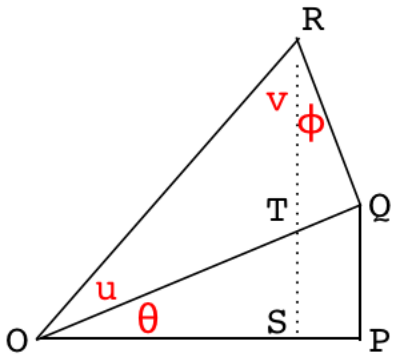
The main result we are headed for is the Pythagorean Theorem. Before we get there, however, it is worthwhile to continue our development of basic geometry with a discussion about right angles and right triangles.

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 not usually do that. The side opposite P, namely c, is the hypotenuse.



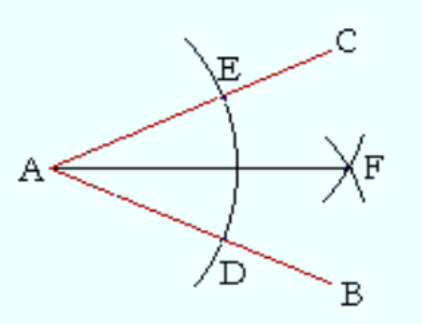
Since the sum of angles in a triangle is equal to two right angles, the sum of the angles θ and ϕ above is also equal to a right angle, or 90 degrees. Angles θ and ϕ are said to be complementary. This fact is often exploited in proofs. Here is an example we will see later on:



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

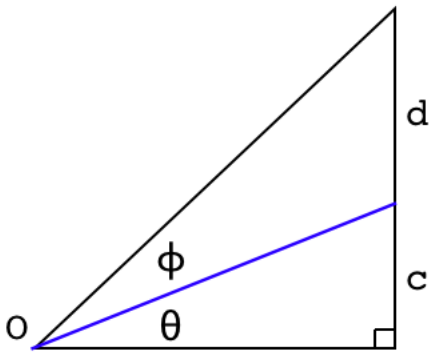
With that background, we now consider a classic problem: involving angle bisectors. Actually, before we do that, let's just show a method for constructing an angle bisector



To bisect angle $\angle BAC$ use the compass to mark off equal segments AD and AE and then mark off equal segments DF and EF. The line segment AF bisects the angle.

Proof: $\triangle ADF$ is congruent to $\triangle AEF$ by SSS. Therefore, $\angle CAF = \angle BAF$.

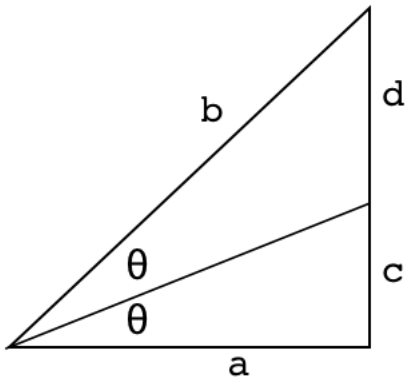
Now, back to our problem, and the diagram below.



Suppose we are given that the large triangle, and the bottom of the two smaller triangles are both right triangles.

We draw a line joining the vertex O on the left with the side opposite. This line could in general be drawn anywhere, however two interesting cases are when the angle at O is bisected, or when the side opposite is bisected.

These cases are different.



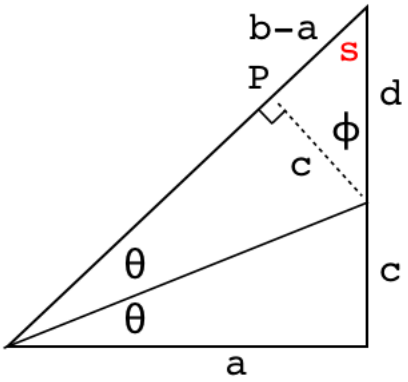
Here we have chosen the first possibility. We are in a position to prove an important theorem.

angle bisector theorem

With reference to the figure above, we are to prove that

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

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



By congruent triangles (the two triangles each with vertex angle θ), the altitude has length c.

By the rules for complementary angles discussed above:

$$2\theta + s = 90 = s + \phi$$

Hence, $2\theta = \phi$. We conclude that the smallest triangle at the top right of the figure is similar to the original. By similar triangles, we form the equal ratios of the hypotenuse to the adjacent angle (either ϕ or 2θ):

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

This is rearranged simply to give

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

which is what we were asked to prove.

□

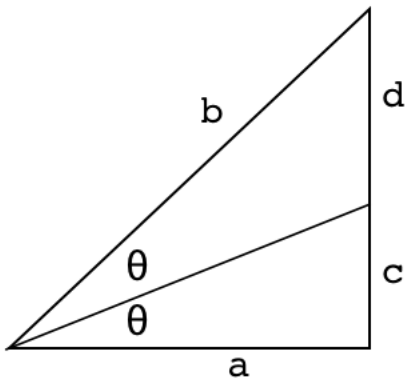
The result can be pushed a little further:

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

Here's the key point

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

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

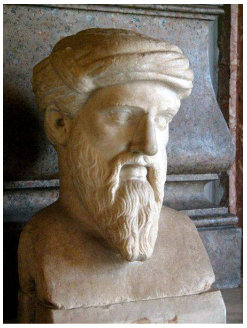


which is a surprising result and becomes important in looking at Archimedes method for approximating the value of π.

Chapter 12

Pythagoras

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

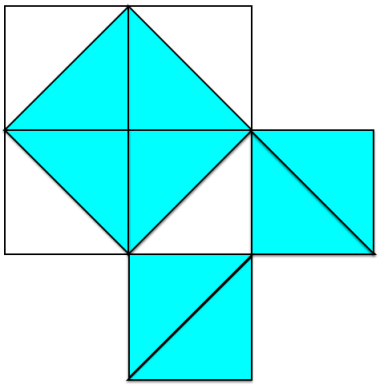


Pythagoras (c.570-c.495 BC) was much younger than Thales but may have encountered him as a young man. Like many other Greek mathematicians, Pythagoras was not from the mainland, but from one of the islands, in his case, Samos, which is not far from Miletus, where Thales lived.

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.

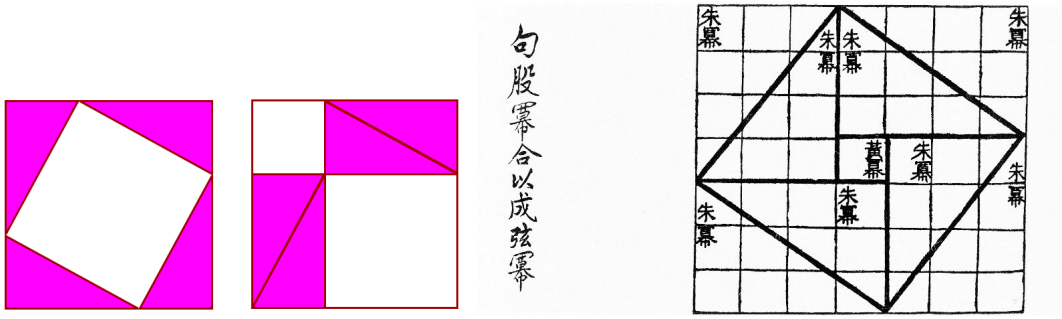


The area of the square on the hypotenuse is equal to twice the area on each side.

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.

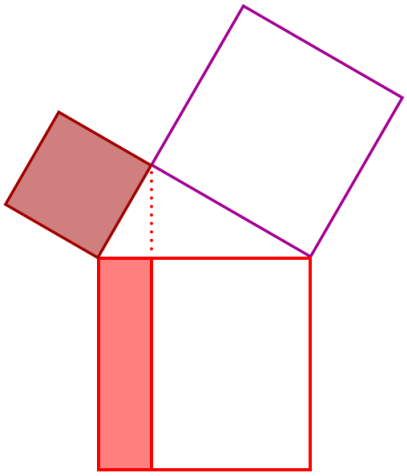
We simply rearrange the triangles. Now we evidently have the same area left over from the four triangles, because they still have the same area and the surrounding box has not changed.

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

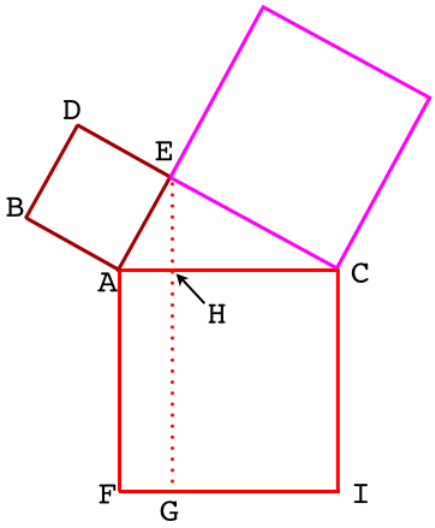
Euclid's proof



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 one of only a few places in the book that we actually show a proof from Euclid, which is a justification for including it.

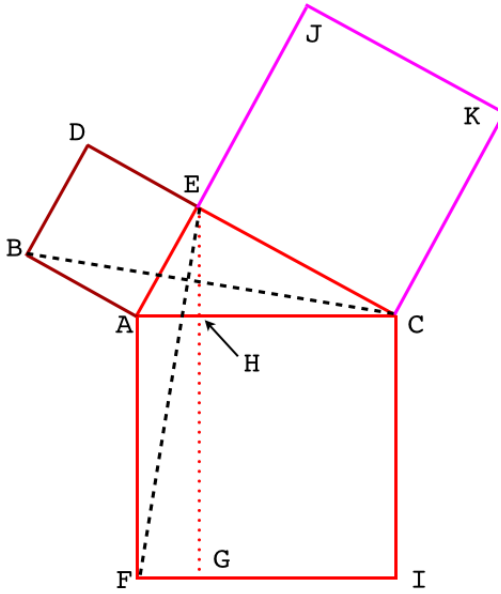
What we will show is that the part of the large square in red is equal in area to the entire small square, in maroon.

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". The two sets of sides are evidently equal

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

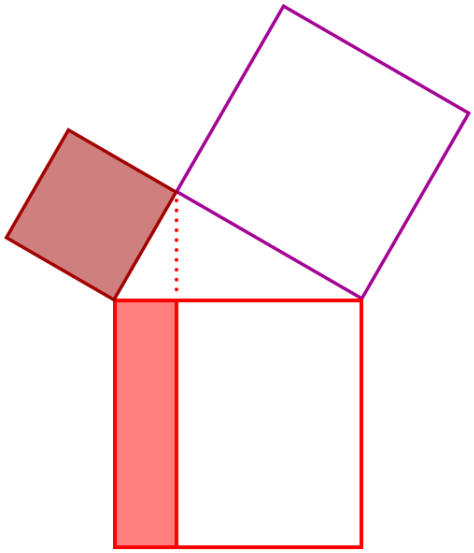
because these are given as sides of two squares.

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

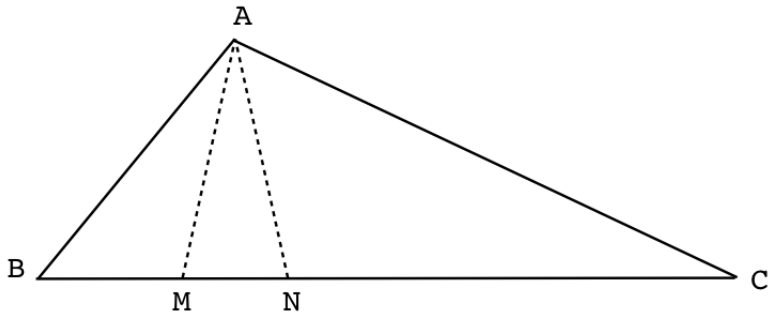


Finally, we could proceed to do the same thing on the right side of the figure, but we just appeal to symmetry. All the equivalent relationships will hold. □

There are a number of easy algebraic proofs of the Pythagorean theorem. These are discussed in a separate chapter **here**.

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

$$AB^2 + AC^2 = BM \times BC + NC \times BC$$

$$= (BM + NC) \times BC$$

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.

Pythagorean triples

The simplest right triangle with integer sides is $3, 4, 5$:

$$3^2 + 4^2 = 5^2$$

any multiple n will work

$$(3n)^2 + (4n)^2 = (5n)^2$$

but that's not so interesting. The triples which are not multiples of another triple are called *primitive*.

To go further, we can use Euclid's formula. For every integer m, n, with $m > n$, a Pythagorean triple is given by

$$a = m^2 - n^2 \quad b = 2mn \quad c = m^2 + n^2$$

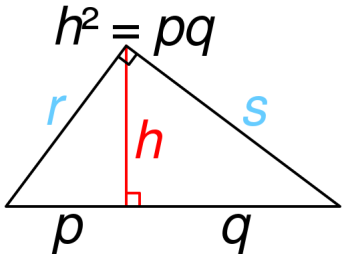
This is discussed further in a separate chapter **here**.

https://en.wikipedia.org/wiki/Pythagorean_triple#Enumeration_of_primitive_Pythagorean_triples

A thousand years before Pythagoras, the Babylonians knew the triple $4601, 4800, 6649$. It seems unlikely that they found this by random search.

geometric mean

As a slight detour from calculus, but on the topic of this chapter



According to the figure, the altitude of a right triangle is related to the two segments along the base by

$$h^2 = pq$$

$$h = \sqrt{pq}$$

That is, h is the geometric mean of these two values p and q.

The proof is simple. 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:

$$2h^2 + p^2 + q^2 = p^2 + 2pq + q^2$$

Part IV

Circles

Chapter 13

Circles

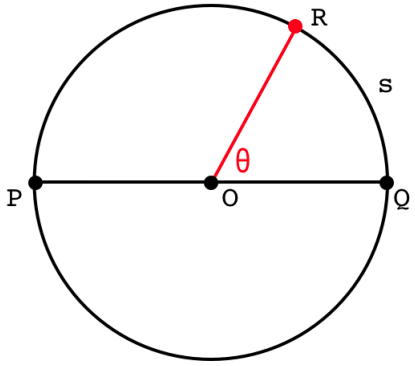
From a previous chapter, Euclid's third postulate was:

∘ 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 called a compass:

https://en.wikipedia.org/wiki/Compass_(drawing_tool)

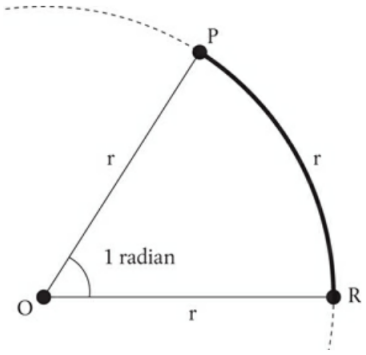
If the radius is extended so that it cuts the circle at two points, it is called a diameter. We saw previously that one can construct a line perpendicular to any given line. If that line is constructed perpendicular to the diameter at the point where it meets the circle, the new line is called a tangent 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. For a unit circle with radius = 1, the total circumference is 2π, so the arc swept out by the angle θ 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.

We say that the angle θ is equal to the arc it sweeps out on the circumference, in radians.

$$s = \theta$$

To convert some more measures of angles in degrees to radians:

$$180^\circ = \pi,\ 90^\circ = \frac{\pi}{2}$$

$$60^\circ = \frac{\pi}{3},\ 45^\circ = \frac{\pi}{4},\ 30^\circ = \frac{\pi}{6}$$

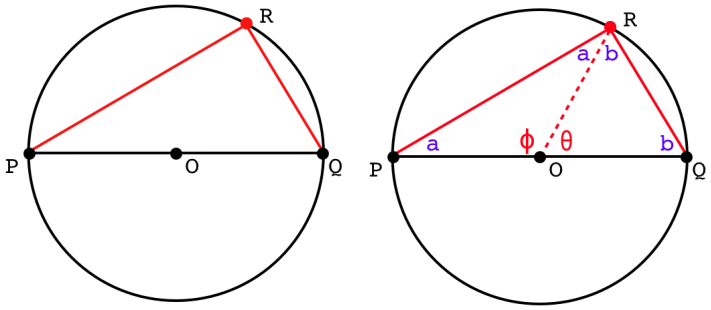
Thales' theorem

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.

Now, 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 $\triangle OPR$ and $\triangle OQR$ are both isosceles.

Label the respective base angles a and b. By considering that together the sum of the angles of $\triangle PQR$ can be written:

$$2a + 2b = \pi$$

$$a + b = \frac{\pi}{2}$$

But this is the measure of $\angle PRQ$.

In addition, the arcs swept out by angles a and b (OPR and OQR on the diameter) clearly add up to π. This suggests that:

$$a = \frac{\theta}{2}$$

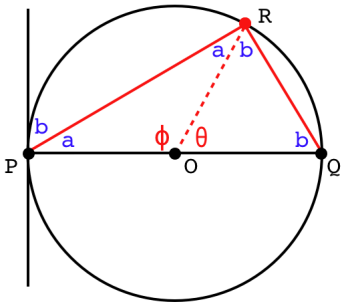
$$b = \frac{\phi}{2}$$

Proof:

$$2a + 2b = \pi = 2a + \phi$$

$$\phi = 2b$$

Consider the chord PR and draw the tangent at P.



The arc between the tangent and the chord equals 2b because it is the same arc as cut off by $\angle\ PQR$ (which is $\angle\ b$).

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):

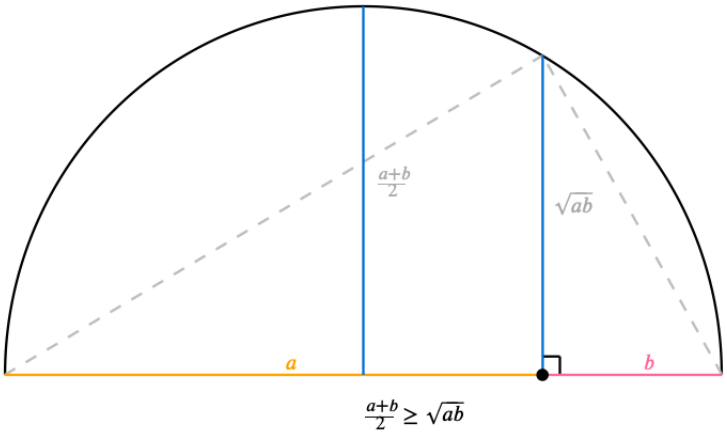
$$m \left[\frac{x_i}{y_i} \right] = \frac{m(x_i)}{m(y_i)}$$

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.

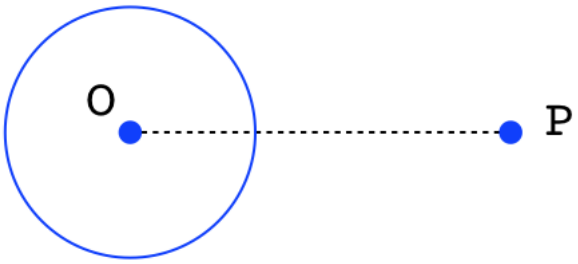
A number of examples are given in the article.

This section is here because of this proof-without-words that the geometric mean is always less than or equal to the arithmetic mean.

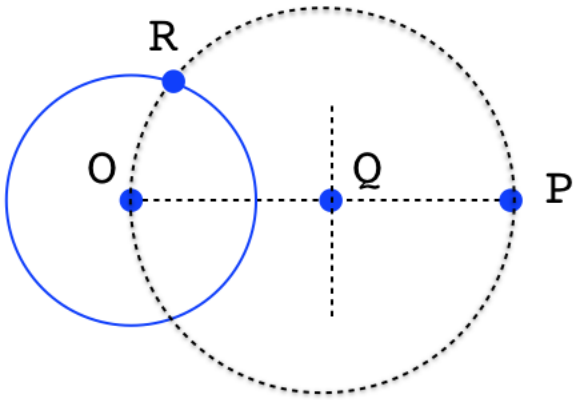


tangents

Thales theorem provides a way to construct the tangent to a circle passing through any exterior point P.

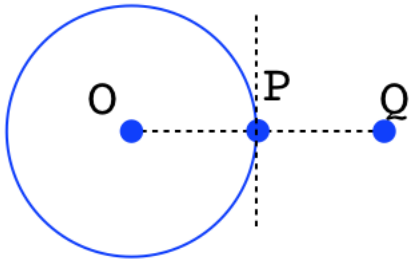


Use OP as the diameter of a circle. Draw the line segment OP and divide it in half by erecting the perpendicular bisector. Use that distance as the radius of a new circle. The point R is the intersection of the two circles.



By Thales theorem, $\angle ORQ$ is a right angle, and since OR is a radius of the original circle, RQ is the tangent at the point R.

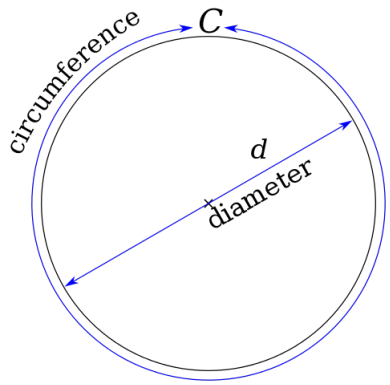
To construct a tangent on a circle at a given point P



Extend OP to Q such that OP is equal to PQ. Construct the perpendicular bisector at P. That is the tangent of the circle.

Chapter 14

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:

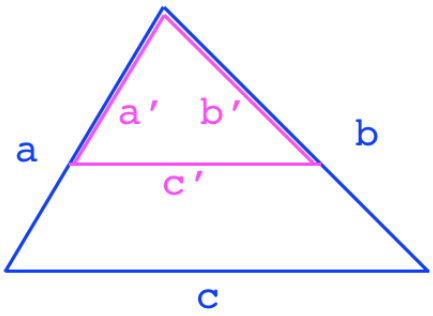
$$\pi = \frac{C}{d} = \frac{C}{2r}$$

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.

Let us first define similarity: two triangles are similar if all their three angles are equal. That is, one is a scaled-down version of the other.

The theorem is: if two triangles are similar, then their sides are proportional to each other.



Draw a horizontal line parallel to the base. Then the resulting small triangle has its sides in proportion to the original one as:

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

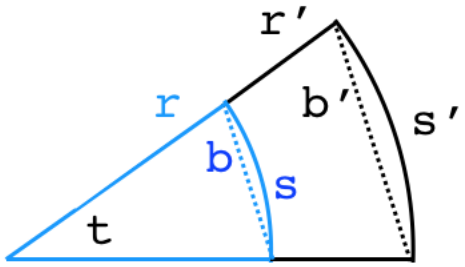
Now we can prove that π is a constant.

Proof:

Consider two circles of different sizes on the same center, with inner radius r and outer radius r'.

Divide the circles into n equal sectors each with central angle t.

Then the arc length $s = C/n$ or $s' = C'/n$. For any particular n we have that $C/s = C'/s'$. Thus, it suffices to show that the ratio $s/r = s'/r'$ (i.e. is constant) to prove the theorem.



In the limit as $n \to \infty$, i.e., as the example sector gets smaller and the total number of sectors gets very large, $b \approx s$ and $b' \approx s'$.

This is just Archimedes' argument, that as the number of sides of an inscribed regular polygon increases without limit, the perimeter of the polygon will be equal to the circumference of the circle. Therefore as $n \to \infty$

$$s = b; \quad s' = b'$$

But the two triangles are similar, because they share the angle t and are both isosceles. Therefore, the ratio b/r is equal to the ratio b'/r' and then

$$\frac{s}{r} = \frac{s'}{r'}$$

so

$$\frac{C}{r} = \frac{C'}{r'}$$

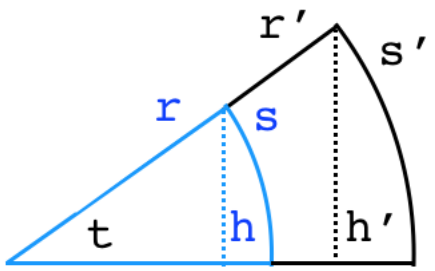
This completes the proof.

□

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 $s = t$, measured in radians.



In the limit that $n \to \infty$, the ratio between s and 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.

□

Pi is irrational

This proof is too challenging for this book. You can read about it in wikipedia, or

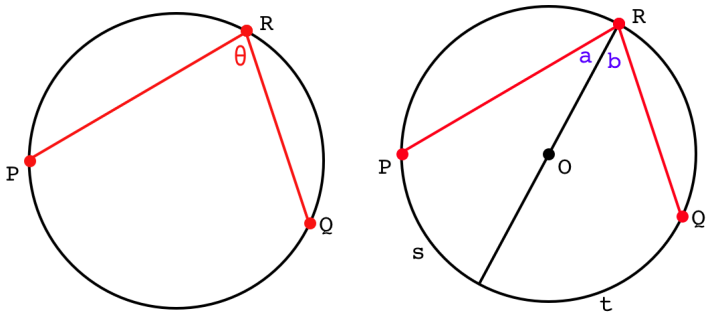
https://mindyourdecisions.com/blog/2013/11/08/proving-pi-is-irrational-a-step-by-step-guide-to-a-simple-proof/

Chapter 15

Arcs of a circle

Having established some basic facts about circles we can do a bit more. We will use some of these results later on.

One is to generalize the result for all arcs. The examples so far contain the diameter in some way. Consider the arc swept out by the angle θ in this figure.



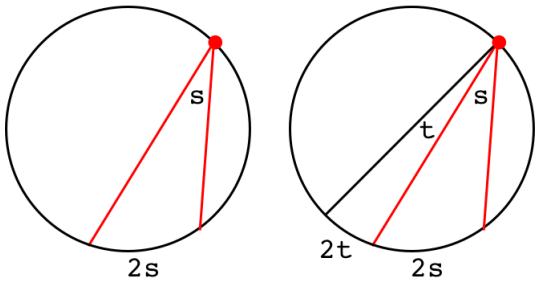
We can prove that the measure of the angle θ is equal to the 1/2 the arc swept out between P and Q. For a simple proof, draw the diameter: By our previous work:

$$b = \frac{t}{2}$$

$$a = \frac{s}{2}$$

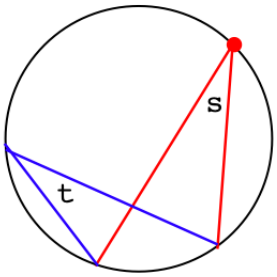
$$\theta = a + b = \frac{s+t}{2}$$

We have proved the theorem for two cases: where the diameter is one line segment flanking the angle, and where the angle includes the diameter. However, the theorem is true even if the angle does not include the diameter.



On the right, draw the diameter. Notice that we have 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 subtraction.

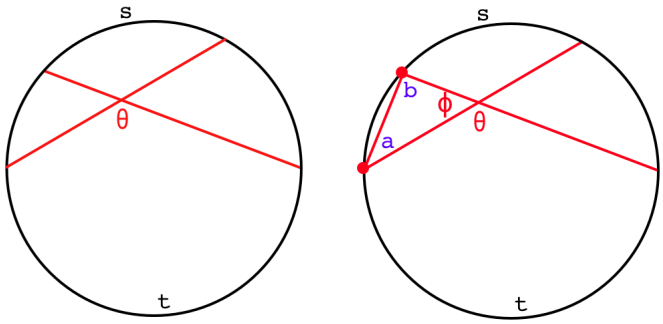
As a corollary, any two angles with vertexes on the circle that cut off the same arc are equal. In the figure, $s = t$. Also the triangles are similar triangles.



Intersecting chords

Given two chords, to prove:

$$\theta = 1/2(s+t)$$

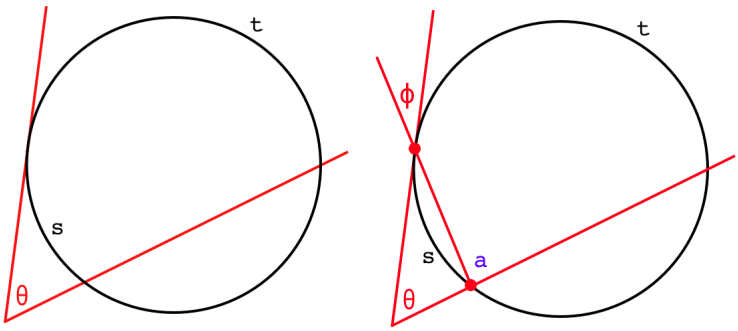


θ is the average of the two arc lengths. Solution: Draw a triangle.

$$a = \frac{s}{2}$$
$$b = \frac{t}{2}$$
$$a + b = \theta = \frac{s+t}{2}$$

Tangent and secant

Rather than having all three points on the circle, one is now outside. We have the same arc swept out by the endpoints (t), but the included angle is now smaller, and there is a new small piece of arc length s.



To prove:

$$\theta = \frac{t-s}{2}$$

Solution: Draw the triangle. By our previous work (and supplementary angles):

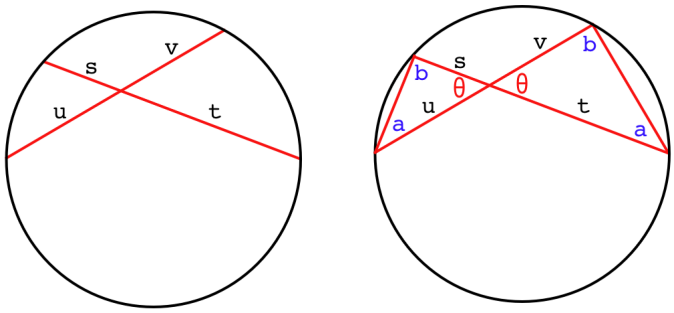
$$\phi = \frac{s}{2}$$
$$a = \frac{t}{2}$$

by supplementary angles:

$$\theta + \phi = a$$
$$\theta = \frac{t}{2} - \frac{s}{2}$$
$$= \frac{t-s}{2}$$

Chord segments

Finally, there is a simple algebraic relationship between chord segments. Draw two chords of the circle and label the lengths of the segments as shown (note: s and t do not refer to arcs any more).



Draw the two triangles. Notice that the two angles labeled a are equal because they sweep out the same arc of the circle, and similarly for the two angles labeled b. By similar triangles:

$$s/u = v/t$$

$$st = uv$$

Chapter 16

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.

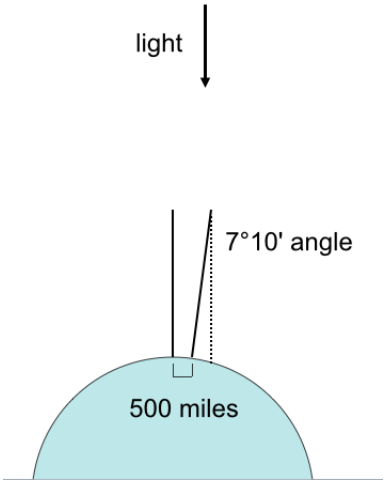
https://en.wikipedia.org/wiki/Eratosthenes

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

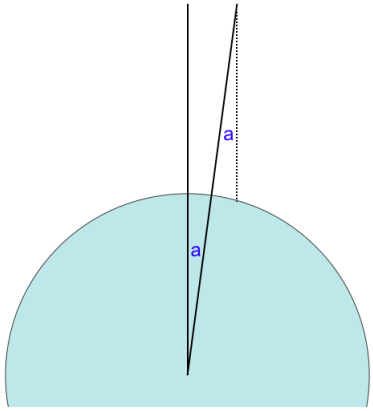
upstream of Luxor, which includes the famous site called the Valley of the Kings. At 24.1 degrees north latitude, Aswan or Syene is close enough to having the sun directly overhead on June 21. (The "Tropic of Cancer" is at 23 degrees, 26 minutes north).

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.

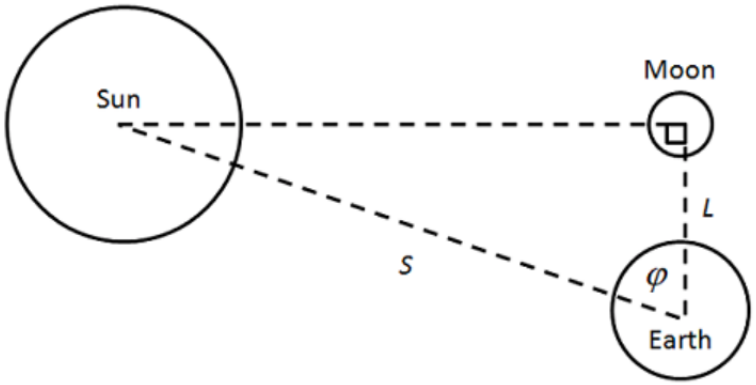
Aristarchus

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 17

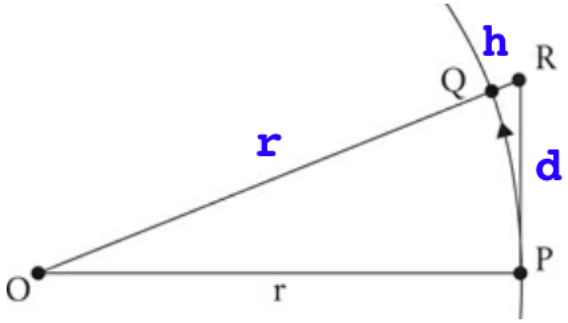
Circular orbits

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:



$$r^2 + d^2 = (r+h)^2 = r^2 + 2rh + h^2$$
$$d^2 = 2rh + h^2$$

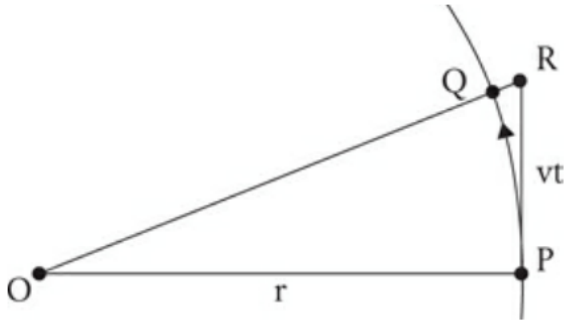
If $h << r$ then we can ignore the very small quantity h^2 and obtain

$$d^2 = 2rh$$
$$r = \frac{d^2}{2h}, \qquad h = \frac{d^2}{2r}$$

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:

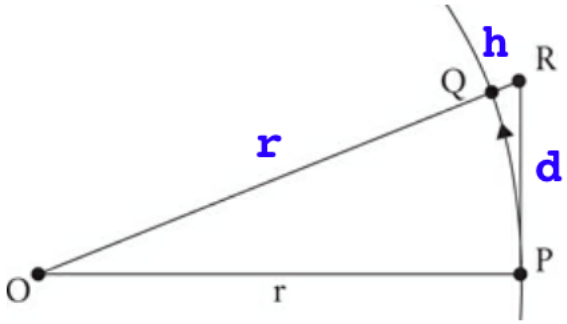
$$d = vt$$

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 $\overline{r}$ 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 V

More number sets

Chapter 18

Rationals

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 $\mathbb{Z}$. 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 is a factor of q or q is a factor of p, then we can divide both top and bottom by whichever is smaller (to yield an integer in the case where $q < p$).

q must not be zero because division by zero is not defined. We *could* choose to allow division by zero, but would quickly run into logical contradictions.

interpolation

Now, consider two rational numbers, not equal. Let

$$s = \frac{p_1}{q_1} \quad t = \frac{p_2}{q_2}$$

Suppose $s < t$.

The *average* of these two rational numbers is:

$$r = \frac{1}{2} \left[s + t \right]$$

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.

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: either $a = b$, $a < b$ or $b < a$.

$$\frac{p}{q} < \frac{s}{t} \iff pt < qs$$

p/q is less than s/t if and only if $pt < qs$. Ordering of the integers guarantees ordering of the rational numbers.

Note: we used the property of the integers that if

$$a < b$$

then 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.

The method for computing the average of two rational numbers could be used to achieve the same thing. The result is:

theorem

• Between *any* two rational numbers it is always possible to find another rational number.

Chapter 19

Euclid's algorithm

The next two chapters don't have a lot to do with calculus, but are here to show some key results of Euclid. One is the Euclidean algorithm, and another is the Fundamental Theorem of Arithmetic. Though he never stated the fundamental theorem in its modern form, Euclid had all the pieces figured out.

It shows some more sophisticated but not too challenging proofs.

This can be skipped without interfering with the development of the rest of the text, except that the fundamental theorem is used in the chapter on irrationals to greatly simplify two proofs about irrational numbers.

Euclidean algorithm

Consider 2 natural numbers a and b. Usually a is allowed to be an integer (i.e., it can be negative), but 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). Here's an example:

```
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
```

First we write the unique prime factorization of a and b (see below). Then 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.

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

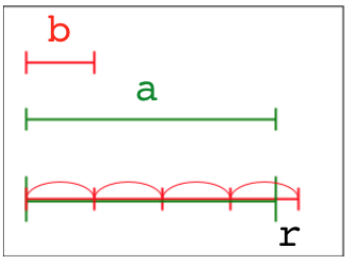
◦ switch $a = b$ and $b = r$ and repeat.

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

$$a = b \cdot q + r$$



Proof:

◦ Either

$$a = b \cdot q$$

◦ Or

$$b \cdot q < a < b \cdot q + b$$

So then

$$a - bq > 0$$

$$a - bq < b$$

and we set $r = a - bq$.

So then let u be the largest integer that divides both a and b

$$a = su$$

$$b = tu$$

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.

There is much more developed in the *extended* Euclidean algorithm that is useful for cryptography.

Chapter 20

Primes

prime numbers

As you know, the positive integers $a > 1$ are of two types. Prime numbers have no factors other than themselves and 1, while composite numbers have at least one other factor. If they are not perfect squares they have two.

The first few primes are:

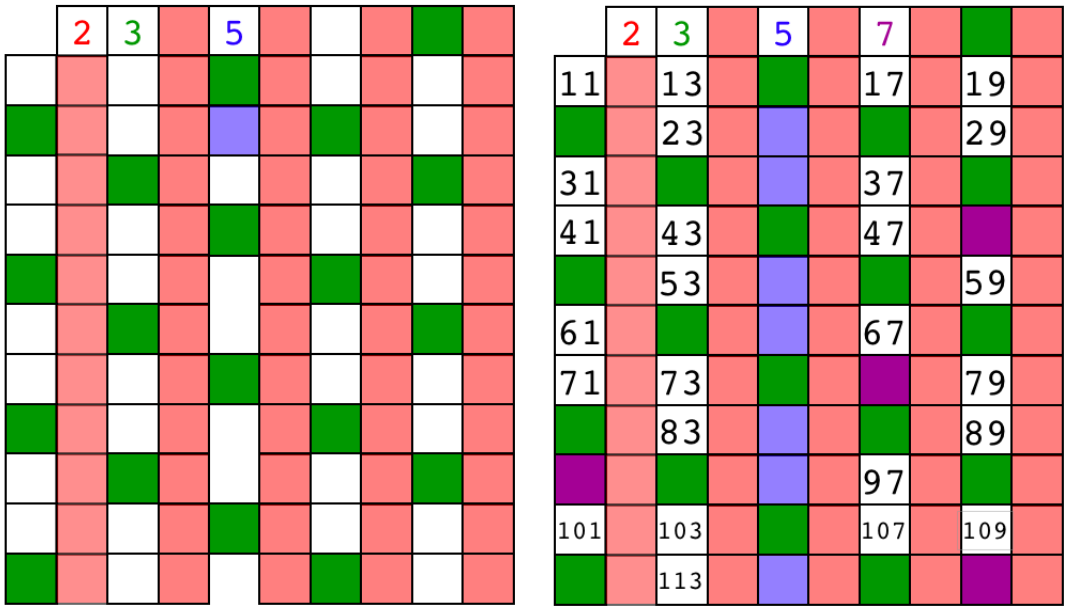
```
2 3 5 7 11 13 17 19 23 29 ...
```

The sieve of Eratosthenes

Eratosthenes is famous in mathematics for his "sieve" which allows one to compute the prime numbers in an economical fashion. We took note of him previously in talking about the circumference of the earth.

The sieve is operated by first enumerating all the integers to some upper limit (here 120). To do things 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.

Starting with the first prime number, 2 (red), eliminate all the numbers divisible by 2 (all the even numbers). Here this has been done by coloring red all of the squares in the even numbered columns (all 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 and 15 go away at this step.

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

The sieve ends when the number for the beginning of the next round, the smallest number not yet eliminated, is greater than the square root of the upper limit (here $\sqrt{120}$). So 7 is used for the last round, 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.

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.

infinite primes

Euclid has a proof that the number of primes (the size of the set of primes) is infinite.

The proof is by contradiction:

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

$$q = (p_1 \times p_2 \times \ldots p_k)$$

$$r = q + 1$$

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

Therefore, none of the known primes divides r, and r is either a prime not in the set of known primes, or the set was originally incomplete.

In either case, the assumption that the set of primes is finite leads to a contradiction.

□

Even for a relatively small number of primes, the second case may hold. Consider

$$(2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13) + 1 = 30031$$

30031 is not prime but is divided by two primes not in the list: 59 and 509.

prime factorization

We will prove that every integer has a unique prime factorization. This is also called *the fundamental theorem of arithmetic*.

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

First, we need a preliminary result, which is called *Euclid's lemma*.

Every natural number $n > 1$, i.e. every positive integer greater than 1, is either prime, or it is the product of two smaller natural numbers a and b.

But the same is true of a and b in turn.

Therefore, every number that can be factored into a and b is the product of the prime factors of a times the prime factors of b.

The notation $m|n$ means m divides n (evenly).

Suppose a given prime p divides $n = ab$, i.e. $p|n$. Then either $p|a$ or $p|b$ (or both).

Proof of existence

The proof is by induction.

Assume the lemma is true for all numbers between 1 and n.

It is certainly true for say, $n \leq 100$, because we can check each case. Start with $n = 101$.

◦ If n is prime (as it is here) there is nothing to prove and we move to $n + 1$.

◦ n is not prime, then there exist integers a and b (with $1 < a \leq b < n$) such that $n = a \times b$.

◦ By the induction hypothesis, since $a < n$ and $b < n$, a has prime factors $p_1 p_2 \ldots$ and b has prime factors $q_1 q_2 \ldots$ so

$$n = ab = p_1 p_2 \ldots q_1 q_2 \ldots$$

This shows that there exists a prime factorization of n.

Proof of uniqueness

To show that the prime factorization is unique suppose that n is the smallest integer for which there exist two different factorizations:

$$n = p_1 p_2 \cdots = q_1 q_2 \ldots$$

Pick the first factor p_1. Since p_1 divides $n = q_1 q_2 \ldots$, by Euclid's lemma, it must divide some particular q_j. Rearrange the q so that q_j is first.

But since p_1 divides q_1 and both are prime, it follows that $p_1 = q_1$.

Now continue the same process with all the factors p_i.

wikipedia:

This can be done for each of the m p_i's, showing that $m \leq n$ and every p_i is some q_j. Applying the same argument with the p and q reversed shows $n \leq m$ (hence $m = n$) and every q_j is a p_i.

□

elegant proof

Hardy and Wright (*Theory of Numbers*, sect. 2:11) have a second proof, which is quite delightful. It is given here almost verbatim:

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

Different factorization:

The same prime P cannot appear in two different factorizations of n, for, if it did, n/P would be abnormal and yet $n/P < n$, the smallest abnormal number.

Thus, we have that

$$= p_1 p_2 \cdots = q_1 q_2 \ldots$$

where the p and q are primes, and no p is a q and no q is a p.

If there exist abnormal numbers with two such factorizations, they must be completely different.

the contradiction

We may take p_1 to be the least p (if the least q is less than the least p, switch labels on all the p's and q's).

Since n is composite, $p_1^2 \leq n$.

The same is true for q_1 and (since $p_1 \neq q_1$), we have that $p_1 q_1 < n$.

Hence, if $N = n - p_1 q_1$, we have $0 < N < n$ and also that N is not abnormal.

Now $p_1 | n$ and since $N = n - p_1 q_1$, so $p_1 | N$.

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

From this 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 $p_2 p_3 \ldots$.

Since q_1 is not a p, this is impossible. Hence there cannot be any abnormal numbers, and this is the fundamental theorem.

□

Chapter 21

Irrationals

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 $p/q = \sqrt{2}$.

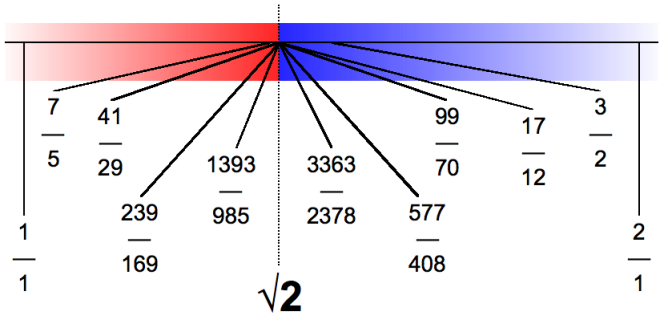
discussion

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

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

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

$$n^2 = p_1^2 \cdot p_2^2 \ldots p_i^2$$

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.

$$\sqrt{2} - 1 = \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ldots}}}$$

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 $\ldots$. 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 = 71/29
2 + 29/71 = 171/71
2 + 71/171 = 413/171
```

To terminate we need to use that initial 1:

```
1 + 171/413 = 584/413 = 1.414043
```

To six places, $\sqrt{2} = 1.414213$. We have only three places, but can get more (convergence is relatively slow, however).

geometric proof

There are many other proofs of the irrationality of the square root of 2.

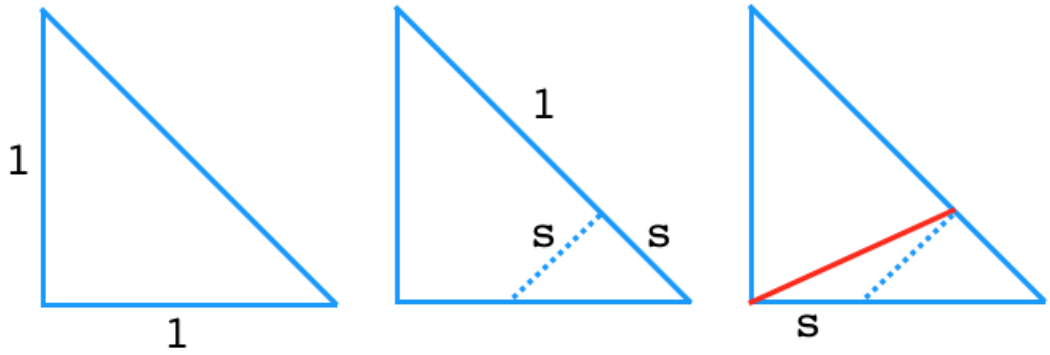
https://www.cut-the-knot.org/proofs/sq_root.shtml

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 the length of the hypotenuse is a rational number, and that its ratio to the side is in "lowest terms".



Now mark off the length of the side on the hypotenuse and erect a perpendicular (middle panel). The new small triangle that is formed is also isosceles (it is a right triangle and it also contains one of the complementary angles of the original right triangle). By hypothesis, its side length s is the difference of two rational numbers, so it is a rational number.

Furthermore, the triangle with the red base and blue sides of length 1 is isosceles (right panel), so by complementary angles the triangle with the red base and one side a dotted line has equal angles at its base and so is isosceles. All the lengths marked s are equal.

Therefore, the hypotenuse of the new, small right triangle is a rational number, since it is equal to $1 - s$.

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 it is discussed **here**. What follows is a much simpler proof based on the fundamental theorem of arithmetic.

We suppose that there exist two integers a and b such that

$$(\frac{a}{b})^2 = n$$

According to the fundamental theorem of arithmetic, both a and b have a unique prime factorization. Suppose that gives $a = a_1 \cdot a_2 \ldots a_i$ and likewise for b so:

$$(\frac{a_1 \cdot a_2 \ldots a_i}{b_1 \cdot b_2 \ldots b_j})^2 = n$$

If every factor b were some a_i, then we could cancel all of them and so a/b would be an integer.

If a/b is to be rational but not an integer, there must be at least one prime factor of b that cannot be cancelled. Call that (those) q, so in lowest terms we have

$$(\frac{a_1 \cdot a_2 \ldots}{q_1 \ldots})^2 = n$$

But then, after squaring, we will have q_1^2 in the denominator and no corresponding factor of either q_1 or q_1^2 in the numerator. Thus, they cannot be canceled and 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.

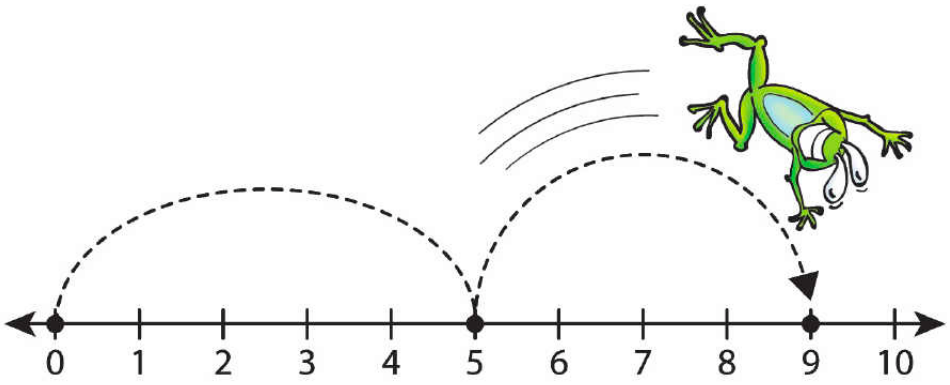
The proof that π is irrational is a bit harder, so we defer that as well.

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 will be given separately **here**.

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 VI

Analytic geometry and trigonometry

Chapter 22

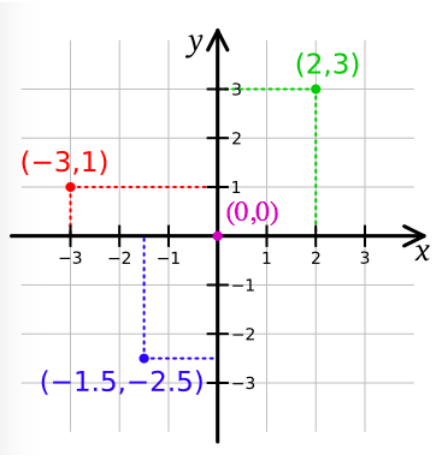
Analytic geometry

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 for-

mula, 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

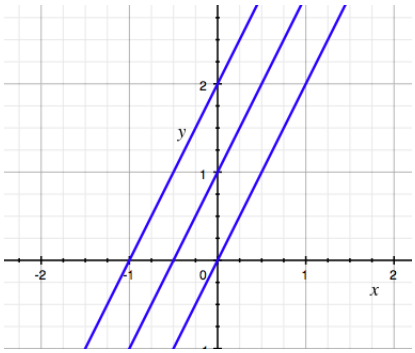
$$y = mx + y_0$$

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:

$$2x - 1 = -x + 8$$

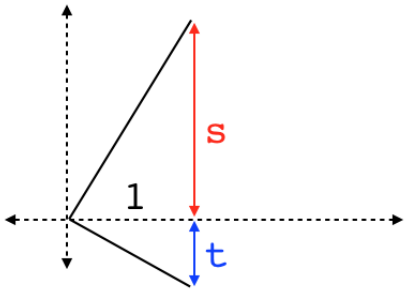
$$3x = 9$$

$$x = 3$$

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 the 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 goes through $(1, t)$.

Recall from the chapter on the Pythagorean theorem that the altitude squared is equal to the product of the two pieces of the base. Here:

$$1^2 = 1 = st$$

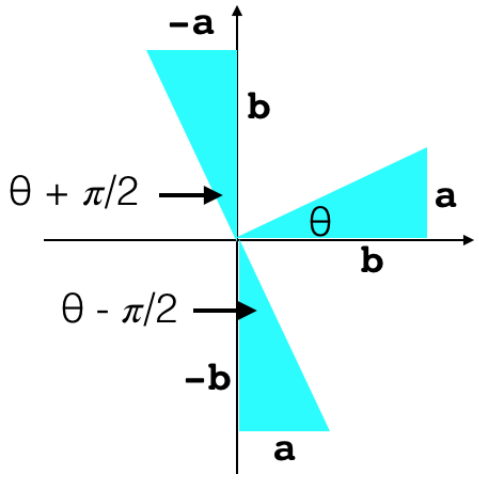
These are the lengths, i.e. the absolute values of the slopes. Thus $|s| = 1/|t|$. But

Clearly the sign of t is negative. So we arrive at

$$s \cdot (-t) = 1$$

$$m_1 = -\frac{1}{m_2}$$

We'll see a natural easy proof of this once we look at trigonometry. Here is a hint:



formula for a circle

A circle can be defined as all the points at the same distance from a central point, let us label that point (h, k). The distance from the points to the center is the radius, denoted r.

Using the Pythagorean theorem, we can calculate the square of the distance from the origin as

$$r^2 = (x - h)^2 + (y - k)^2$$

The simplest circles are those whose central point is the origin of the coordinate system. In that case the equation simplifies to

$$r^2 = x^2 + y^2$$

Usually, we know the value of r and we want to write an equation for y in terms of x. Then

$$y^2 = r^2 - x^2$$
$$y = \sqrt{r^2 - x^2}$$

formula for a parabola

A general formula for a parabola with its vertex at the point (h, k) is

$$y - k = a(x - h)^2$$

where a is called the shape factor. It governs how steeply the curve rises (and by its sign, in which direction it opens).

Multiplying out:

$$y - k = a(x^2 - 2xh + h^2)$$
$$y = ax^2 - 2ahx + ah^2 + k$$

In this form the cofactors are usually simplified as

$$y = ax^2 + bx + c$$

where

$$b = -2ah; \qquad c = ah^2 + k$$

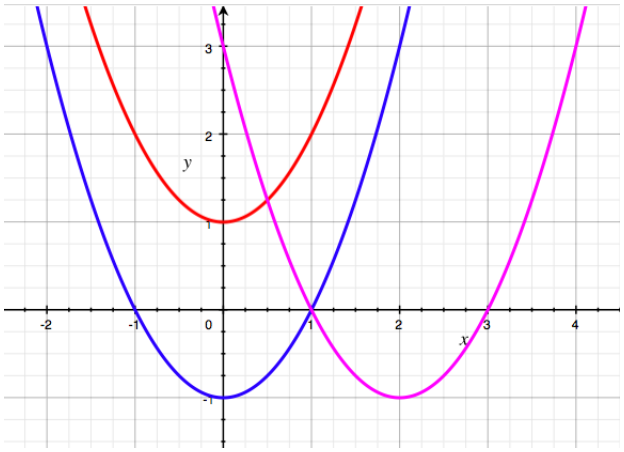
If the equation is given in the second form then we can find:

$$h = -\frac{b}{2a}$$

$$k = c - ah^2 = c - \frac{b^2}{4a}$$

Probably the most common thing we're asked to do with a quadratic equation like this is to find the roots, the values of x for which $y = 0$ is a solution. These are the points where the graph of the curve crosses the x-axis.

It is possible to have 0, 1 or 2 roots.



In the figure, the red curve does not cross the y-axis. Its equation is $y = x^2 + 1$, and there are no solutions, no (real) values of x that solve the equation when $y = 0$.

$$0 = x^2 + 1$$

$$x^2 = -1$$

To find the roots of

$$ax^2 + bx + c = 0$$

We can guess solutions by trying to factor into a form like:

$$(x - s)(x - t) = 0$$

The case of a single root occurs when $s = t$ so we have $a(x - s)^2 = 0$. The common example of that is a parabola with its vertex at the origin, $y = ax^2$.

Roots do not have to be integers (or even rational). An arguably more productive and certainly more general approach is the process of *completing the square.*

First, multiply through by $1/a$ and rearrange:

$$x^2 + \frac{b}{a}x = -\frac{c}{a}$$

The key insight is to recognize that if we add $(b/2a)^2$ to both sides, the left-hand side will become a perfect square:

$$x^2 + \frac{b}{a}x + (\frac{b}{2a})^2 = -\frac{c}{a} + (\frac{b}{2a})^2$$

$$(x + \frac{b}{2a})^2 = -\frac{c}{a} + (\frac{b}{2a})^2$$

$$x + \frac{b}{2a} = \pm\sqrt{-\frac{c}{a} + (\frac{b}{2a})^2}$$

Multiplying top and bottom of the first term under the square root gives a common factor:

$$x + \frac{b}{2a} = \pm\sqrt{-\frac{4ac}{4a^2} + (\frac{b}{2a})^2}$$

which can come out of the square root and then matches what's in the second term on the left-hand side:

$$x + \frac{b}{2a} = \pm\frac{\sqrt{-4ac + b^2}}{2a}$$

which we rearrange slightly to give the standard *quadratic formula*:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

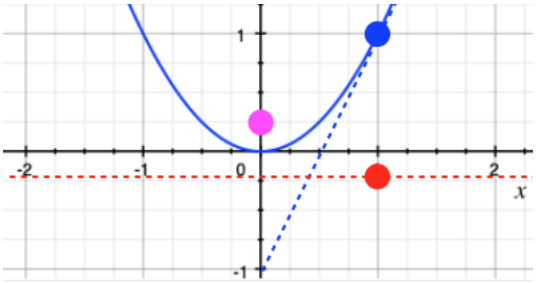
focus and directrix

There is also a classic geometric definition of the parabola.

Based on what we said above, we can transform any parabola of the form $y = ax^2 + bx + c$ into a $(y - k) = a(x - h)^2$. If we're interested in the shape of the parabola and don't care about its absolute location, then without loss of generality, we can translate any parabola to the origin of coordinates, with equation $y = ax^2$, so let us just work with that.

Now, pick a point on the y-axis a distance p up from the origin, colored magenta in the figure. This point is called the focus.

Then draw a line parallel to the x-axis which intersects the y-axis the same distance p below the origin. This line is called the directrix. It is colored red and is dashed.



The parabola consists of all those points whose distance to the focus is equal to the vertical distance to the directrix.

Pick an arbitrary point on the parabola (in blue), with coordinates (x, ax^2). The squared distance to the focus (magenta) is

$$x^2 + (ax^2 - p)^2$$

where Δx is just equal to x and Δy is equal to $y - p$, with $y = ax^2$.

The squared distance to the directrix (red) is $(ax^2 + p)^2$.

For the correct choice of p these distances will be equal:

$$x^2 + (ax^2 - p)^2 = (ax^2 + p)^2$$

We have $(m-n)^2$ on the left-hand side and $(m+n)^2$ on the right-hand side, so the result will have $4mn$ on the right hand side:

$$x^2 = 4apx^2$$

Divide by x^2

$$1 = 4ap$$

$$p = \frac{1}{4a}$$

The shape factor a determines the distance of the focus from the origin, we label that distance as p. The equation of the directrix is $y = -p$.

slope of the tangent

It will turn out that the slope of the tangent to $y = ax^2$ at any fixed point x is equal to $2ax$.

This is literally the first result from differential calculus, but we will also see a way to find it using analytical geometry in the next chapter, as well as a vector approach later on.

Thus, the equation of a line passing through the point (x, ax^2) with the given slope is

$$y' - ax^2 = 2ax(x' - x)$$

where (x', y') is any other point on the line.

What *that* means is that the x-intercept x_0 of the tangent line is:

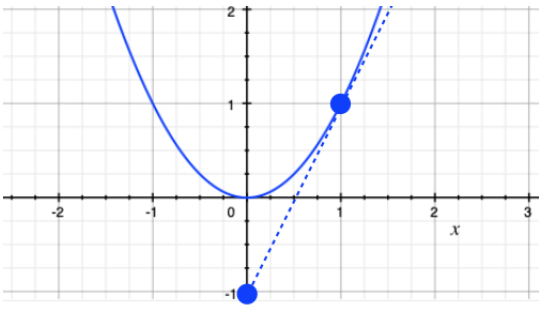
$$-ax^2 = 2axx_0 - 2ax^2$$

$$ax^2 = 2axx_0$$

$$x = 2x_0$$

$$x_0 = \frac{1}{2}x$$

The tangent line passes through the x axis halfway back toward the origin.

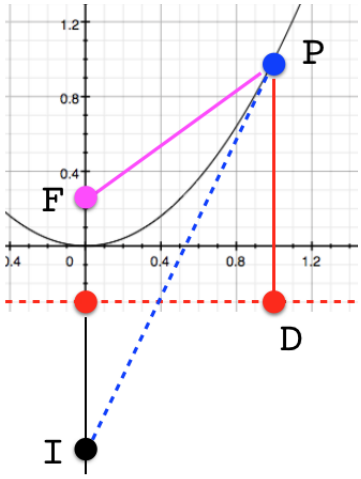


And what *that* means is that the y-intercept is symmetrical with the original point (as far below the x-axis as the point is above it). Here's the algebra:

$$y_0 - ax^2 = 2ax(0 - x)$$

$$y_0 = -ax^2$$

And then finally, if the point on the parabola is P, the focus F, the intersection with the directrix D, and the y-intercept I



the quadrilateral $FPDI$ is a regular parallelogram with all four equal sides, and its long diagonal (the tangent line) makes equal angles with

FP and PD.

And what *that* means is that if PD is extended vertically, the angle it makes with the tangent line is equal to the angle between FP and the tangent line, so that for example, all vertical light rays entering a parabola will reflect and then come together at the focus.

Chapter 23

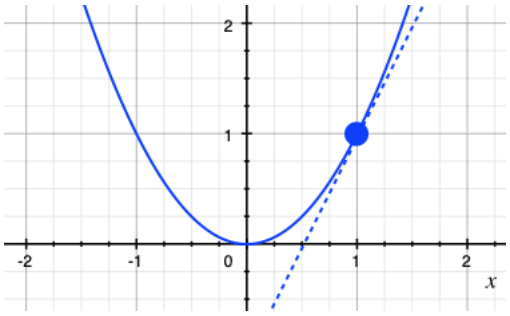
Slope of a parabola

In this chapter, we show how to find the slope of parabola at any point using classical methods.

part 1

Consider the simplest parabola: $y = x^2$.

The point $(1, 1)$ is on the curve, because $(x = 1, y = 1)$ satisfies the equation $y = x^2$.



Suppose we know that the slope of the tangent to the curve at the point $(1, 1)$ is equal to 2.

(Using calculus to find this result is trivial, we'll also show a non-

calculus method in part three, below).

The equation of the tangent line is

$$y_2 - y_1 = m(x_2 - x_1)$$

Plugging in for $(x_2, x_1) = (1, 1)$ (and just writing (x, y) for (x_1, y_1)):

$$y - 1 = 2(x - 1)$$

$$y = 2x - 1$$

Now suppose that we knew only the parabola and this slope, but we did not know the point where the tangent meets the curve, and so do not know the y-intercept.

We have the equation of a line:

$$y = 2x + y_0$$

We seek points which are simultaneously on the line and the curve. They must satisfy both equations.

Since this is a tangent line, we seek the value for which this expression has only a single solution. The tangent "kisses" the curve at a single point.

So, substitute for y from the equation for the curve:

$$x^2 = 2x + y_0$$

$$x^2 - 2x - y_0 = 0$$

Now look at the quadratic formula we would use to solve this equation for x:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

There is a single solution when the part under the square root (called the discriminant) is equal to zero.

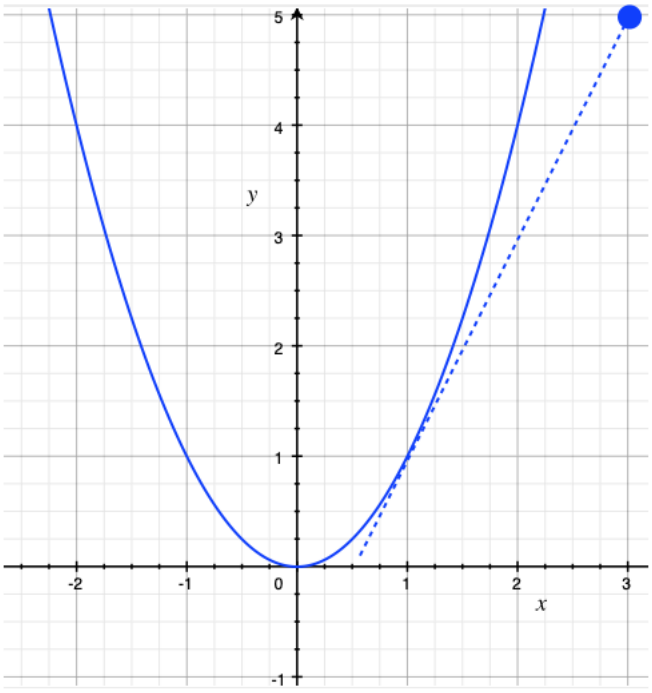
$$b^2 - 4ac = 0$$
$$b^2 = 4ac$$
$$(-2)^2 = 4(-y_0)$$
$$y_0 = -1$$

Therefore, the equation of the tangent line is $y = 2x - 1$, which matches what we had before.

In general, $y = 2x + y_0$ is a *family* of lines. For $y_0 = -1$, there is a single solution for x to be both on the line and the parabola. For $y_0 < -1$, there are no solutions, while for $y_0 > -1$ there are two solutions, because the line actually traces out a secant of the parabola, passing through the curve at two points.

part 2

Now suppose we have the same parabola and a point not on the parabola, but in the plane and outside of the "cup" of the parabola, such as $(3, 5)$. We seek the equations of tangent lines to the parabola that go through this point.



There will be two of them. We show just one in the figure.

The equations of lines passing through this point, with different slopes m are given by:

$$(y_2 - y_1) = m(x_2 - x_1)$$

Here, let (x_2, y_2) be $(3, 5)$ and then multiply by -1, and drop the subscript, to obtain:

$$y - 5 = m(x - 3)$$

Since values of (x, y) are both on the line and the parabola $y = x^2$, we can plug in for y:

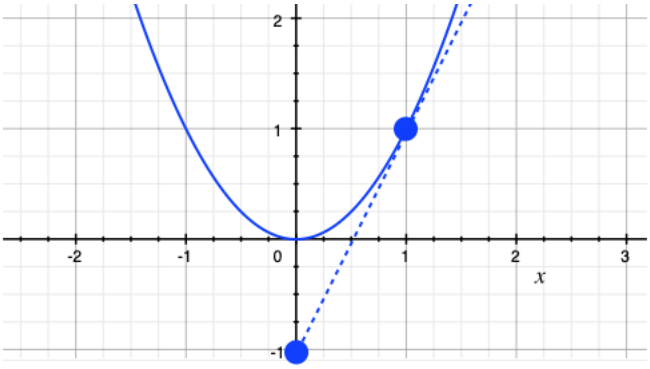
$$x^2 - 5 = mx - 3m$$

$$x^2 - mx + (3m - 5) = 0$$

As before, solutions are given by the quadratic equation. The value of the slope m giving a single solution (zero discriminant) is:

$$(-m)^2 - 4(3m - 5) = 0$$
$$m^2 - 12m + 20 = 0$$
$$(m - 2)(m - 10) = 0$$
$$m = 2, \quad m = 10$$

We knew the first one already, because the point $(3, 5)$ is on the line $y = 2x - 1$. This is the tangent to the curve at $(1, 1)$, which has slope $m = 2$.



Actually, there is always another solution. Any vertical line (with infinite slope) passes through only a single point on the parabola.

Basically what this amounts to is that in the equation

$$x = \frac{m \pm \sqrt{(-m)^2 - 4(3m - 5)}}{2}$$

as m gets very large, only the term $(-m)^2$ matters under the square root, so we have

$$x = \frac{m \pm \sqrt{(-m)^2}}{2}$$

if we choose the negative root, then as $m \to \infty$, $m - \sqrt{m^2} \to 0$.

part 3

Now suppose we are given the same parabola again and also a point on it such as (x_1, y_1).

Any line through that point has the equation:

$$y - y_1 = m(x - x_1)$$

To find the equation of a tangent line through that point we need the slope m.

If there is a point (x, y) that is on the line and *also* on the parabola, it must satisfy $y = ax^2$ as well, so:

$$ax^2 - ax_1^2 = m(x - x_1)$$

$$ax^2 - mx - ax_1^2 + mx_1 = 0$$

Certainly $x = x_1$ is a solution.

The value of m must be such that there are *no other solutions.*

Write the quadratic equation to solve for x:

$$x = \frac{m \pm \sqrt{m^2 - 4a(mx_1 - ax_1^2)}}{2a}$$

There is a single solution when the discriminant is zero, that is, when

$$x = \frac{m}{2a}$$

$$m = 2ax$$

Since $x = x_1$ for the tangent line

$$m = 2ax_1$$

as expected.

The slope of the tangent line is $2ax_1$ and in particular, at the point $(1, 1)$, the slope is equal to 2.

That's the answer, but there are two points to follow up on. We should plug the answer into these two equations and check what happens. We need

$$ax^2 - mx - ax_1^2 + mx_1 = 0$$

and

$$m^2 - 4(mx_1 - ax_1^2) = 0$$

For the first one:

$$ax^2 - 2ax^2 - ax_1^2 + 2ax_1$$

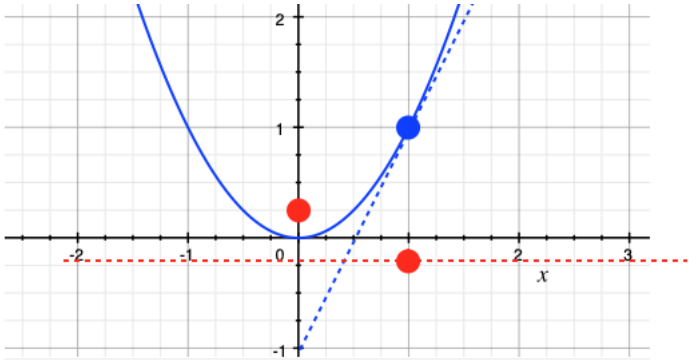
This is certainly equal to zero when $x = x_1$.

Then

$$m^2 - 4a(mx_1 - ax_1^2)$$
$$4a^2x^2 - 4a(2ax_1 - ax_1^2)$$
$$4a^2x^2 - 8a^2x_1^2 + 4a^2x_1^2$$

is also equal to zero when $x = x_1$.

alternate solution

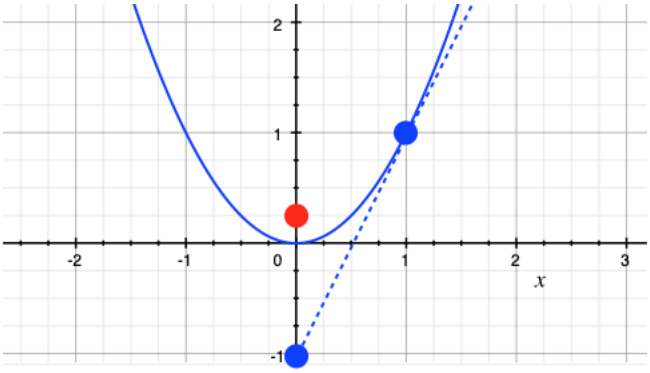


A parabola is defined geometrically by its focus, which is the point $(p, 0)$ for a centered parabola.

The focus is paired with a directrix, which is the line $y = -p$ for a vertex at the origin.

All points on the parabola lie at the same distance d from the focus and the directrix.

A relatively advanced fact about the parabola is that any tangent line intersects the y-axis at the same distance d from the focus.



Which is to say that if we draw a triangle in the above diagram using the two blue points and one red one, the two blue points are the vertices of equal angles and the triangle formed is isosceles.

For $y = x^2$, consider the point (x, x^2), and find the distance to the focus squared as

$$d^2 = (x)^2 + (x^2 - p)^2$$

$$d^2 = x^2 + x^4 - 2x^2p + p^2$$

Call the y-intercept k so then

$$k + d = p$$

$$d^2 = p^2 - 2pk + k^2$$

Equating the two expressions:

$$p^2 - 2pk + k^2 = x^2 + x^4 - 2x^2p + p^2$$

$$k^2 - 2pk = x^2(1 + x^2 - 2p)$$

In this case, we know $x = 1$ and $p = 1/4$ so

$$k^2 - \frac{k}{2} - (2 - \frac{1}{2}) = 0$$

We factor to obtain:

$$(k + 1)(k - \frac{3}{2}) = 0$$

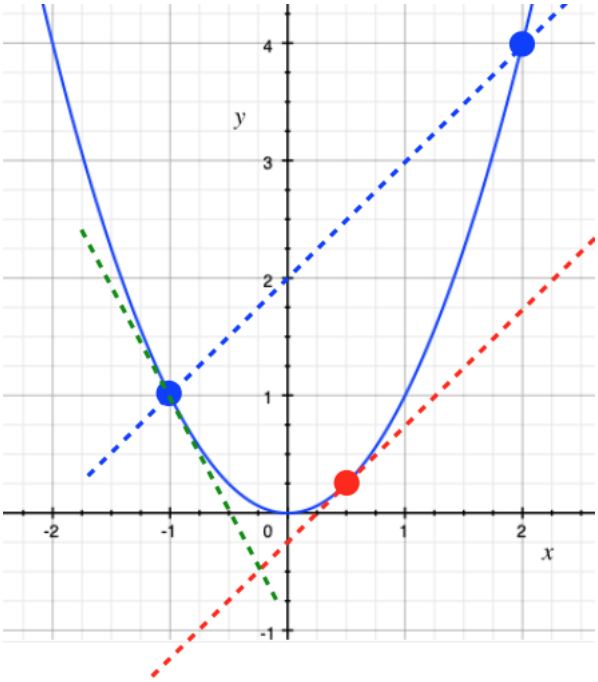
$k = -1$ was our solution above.

I am a little uncertain as to the significance of the other solution ($k = 3/2$). But it cannot be an accident that this is the y-intercept of the line perpendicular to the tangent that goes through the point of tangency.

further comment

The slope of the parabola has some simple interesting properties. For example, pick any two points (x, y) and (x', y') on our standard parabola.

The slope of the line that connects those two points is equal to the slope of the parabola at the point whose x-value is halfway in between.



For the first part:

$$\begin{aligned} m &= \frac{y' - y}{x' - x} \\ &= \frac{ax'^2 - ax^2}{x' - x} \\ &= a\left[\frac{x'^2 - x^2}{x' - x}\right] \\ &= a(x' + x) \end{aligned}$$

For the midpoint

$$x_m = \frac{1}{2}(x' + x)$$

and the slope is

$$\begin{aligned} &2a \cdot \frac{1}{2}(x' + x) \\ &= a(x' + x) \end{aligned}$$

A similar result is that if we pick any two points (x, y) and (x', y'), and draw their slopes, the point where the two slope lines meet has its x-value exactly halfway in between x and x'.

circle

Suppose we have a unit circle and an external point (x, y). We wish to find the equation of the tangent line to a point on the circle. Call that point (a, b).

Circles are special. Any tangent is perpendicular to the radius at the point of tangency.

The line through (a, b) and the origin has slope b/a since

$$m = \frac{b - 0}{a - 0}$$

If we have two lines with slopes m_1 and m_2 and they are perpendicular, the product is -1. So the tangent to the circle at (a, b) has slope $-a/b$ and the line through (x, y) and (a, b) is

$$-\frac{a}{b} = \frac{y - b}{x - a}$$
$$-ax + a^2 = by - b^2$$

We also have that $a^2 + b^2 = 1$ so

$$ax + by = 1$$

Substitute into the equation of the circle:

$$a^2 + (\frac{1 - ax}{y})^2 = 1$$

$$a^2y + 1 - 2ax + a^2x^2 = y$$
$$(y + x^2)a^2 - 2xa + (1 - y) = 0$$

We have a quadratic in a. For a particular x and y, we can solve for a.

ellipse

I found a problem on the web that extends this to the ellipse:

https://math.stackexchange.com/questions/834392/equations-of-lines-tangent-to-an-ellipse

In working that problem, I ended up with a quartic equation (fourth power). This is, quite literally, a mess.

Here's a great idea for ellipse problems: Stretch and rescale the problem to one involving a circle, by using a *change of variable*. Suppose the ellipse is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Let $x = au$ and $y = bv$. Then

$$\frac{a^2u^2}{a^2} + \frac{b^2v^2}{b^2} = 1$$
$$u^2 + v^2 = 1$$

The ellipse has become a unit circle!

Apply the same transformation to any points in the problem, solve the problem, and then reverse the transformation.

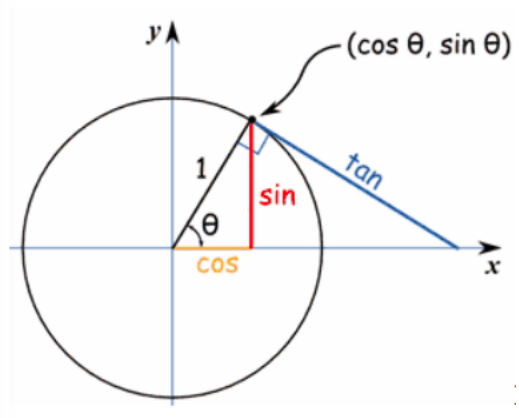
Chapter 24

Six functions

The most elementary trigonometric functions are sine and cosine.

basic definitions

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 diagram you can see that any point (x, y) on the unit circle can be described in radial coordinates as $(\cos\theta, \sin\theta)$.



That is:

$$x = \cos\theta \qquad y = \sin\theta$$

If the circle has radius r then

$$x = r\cos\theta \qquad y = r\sin\theta$$

The tangent is

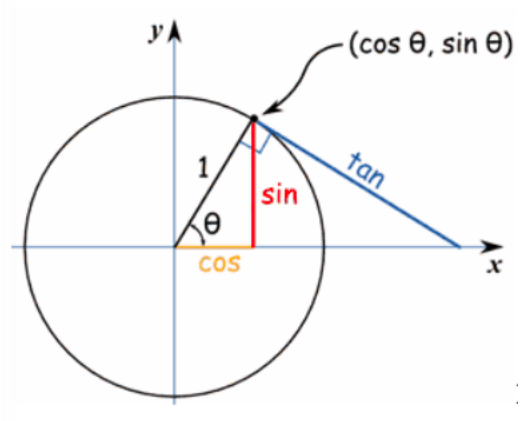
$$\tan\theta = \frac{\sin\theta}{\cos\theta}$$

In the diagram, all three right triangles are similar.

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 it is.



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, but the other two are not especially important in calculus. However, they do come up in one context that we will look at, 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

$$\sin^2\theta + \cos^2\theta = 1$$

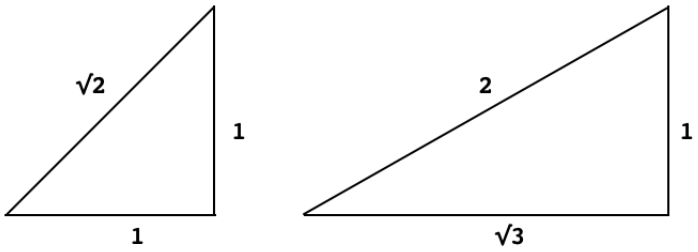
and transformed to

$$1 + \tan^2\theta = \sec^2\theta$$

It is assumed you've studied trigonometry before.

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 $\pi/3$.

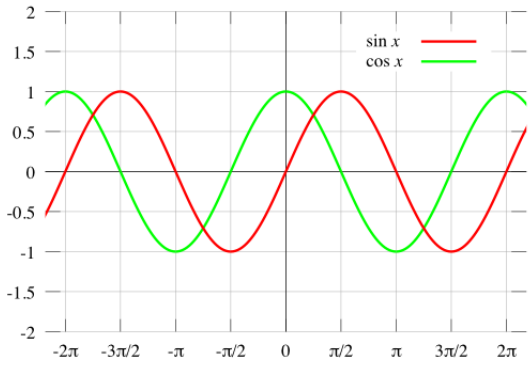
The values are

$$\sin\frac{\pi}{6} = \frac{1}{2} = \cos\frac{\pi}{3}$$

$$\cos\frac{\pi}{6} = \frac{\sqrt{3}}{2} = \sin\frac{\pi}{3}$$

$$\tan\frac{\pi}{6} = \frac{1}{\sqrt{3}}$$

graph

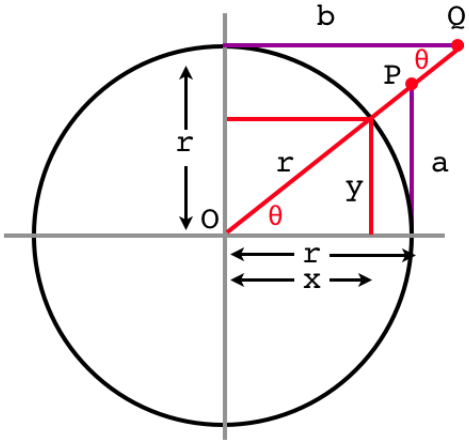


Savov:

> 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.

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

$$r/b = y/x = \tan\theta$$

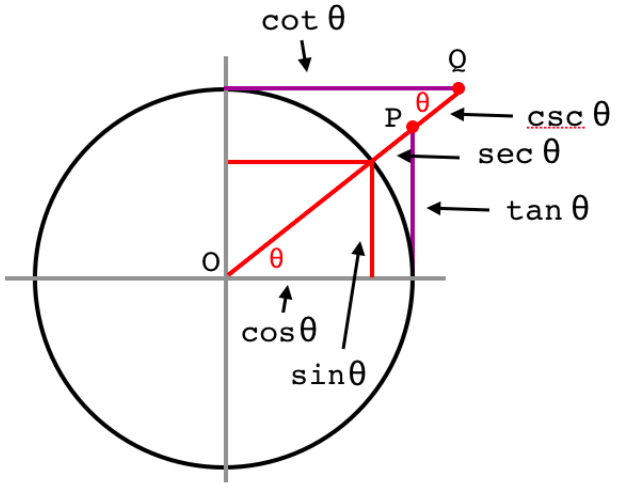
since $r = 1$

$$b = \frac{r}{\tan\theta} = \frac{1}{\tan\theta} = \cot\theta$$

Finally

$$r/OQ = 1/OQ = \sin\theta$$

$$OQ = \frac{1}{\sin\theta} = \csc\theta$$


cot θ
Q
θ
P
csc θ
sec θ
tan θ
O
θ
cos θ
sin θ

Chapter 25

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

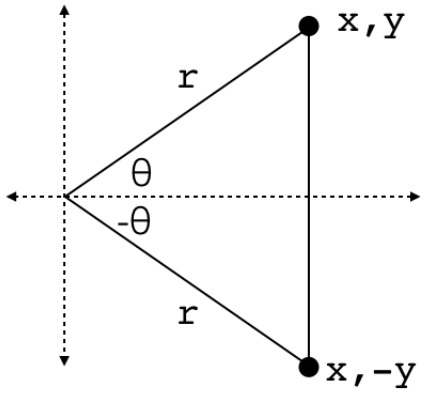
For $\cos s + t$ flip the sign on the second term.

$$\cos s + t = \cos s \cos t - \sin s \sin t$$

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.

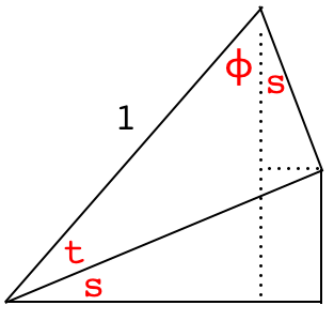
For $\sin s - t$, flip the sign on the second term, as before.

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.

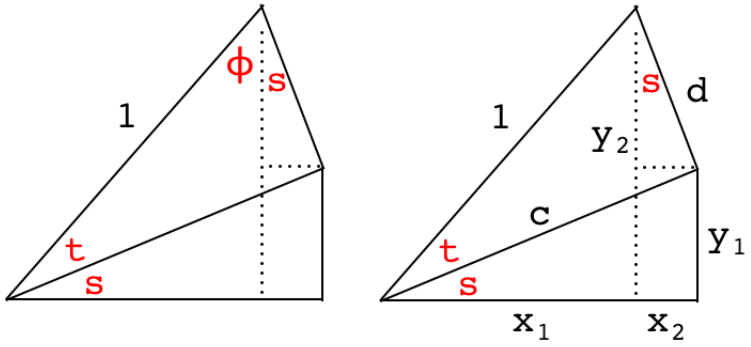
Consider a right triangle, with one of the angles labeled s. Construct another right triangle containing angle t, and scale it so that the base adjacent to angle t is just as long as the hypotenuse of the triangle containing angle s, and draw them one on top of the other as shown:

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

$$\cos s + t = x_1$$

□

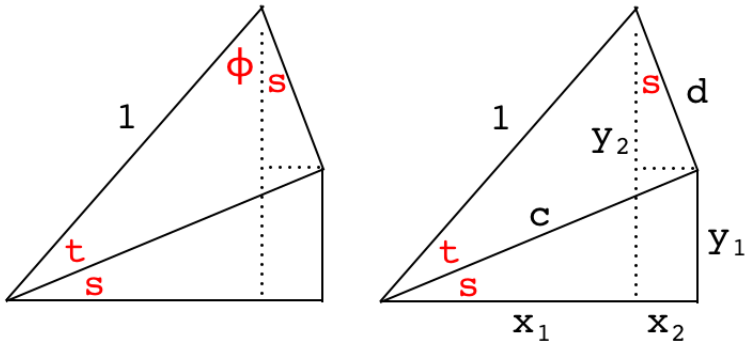
As a quick check we can ask what happens to the formula

$$\cos s + t = \cos s \cos t - \sin s \sin t$$

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)



$$\sin s = \frac{y_1}{c}; \quad \cos t = \frac{c}{1}; \quad \sin s \cos t = y_1$$

$$\sin t = \frac{d}{1}; \qquad \cos s = \frac{y_2}{d}; \qquad \sin t \cos s = y_2$$

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.

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

$$\sin \frac{\pi}{12} = \sin s - t = \sin s \cos t - \sin t \cos s$$
$$= \frac{1}{\sqrt{2}} \cdot \frac{\sqrt{3}}{2} - \frac{1}{2} \cdot \frac{1}{\sqrt{2}} = \frac{\sqrt{3} - 1}{2\sqrt{2}}$$
$$\cos \frac{\pi}{12} = \cos s - t = \cos s \cos t + \sin s \sin t$$

$$= \frac{\sqrt{3}}{2} \cdot \frac{1}{\sqrt{2}} - \frac{1}{2} \cdot \frac{1}{\sqrt{2}} = \frac{\sqrt{3}+1}{2\sqrt{2}}$$

We just check that $\sin^2\theta + \cos^2\theta = 1$:

$$\frac{(\sqrt{3}-1)^2 + (\sqrt{3}+1)^2}{(2\sqrt{2})^2}$$

$$= \frac{3 - 2\sqrt{3} + 1 + 3 + 2\sqrt{3} + 1}{8} = 1$$

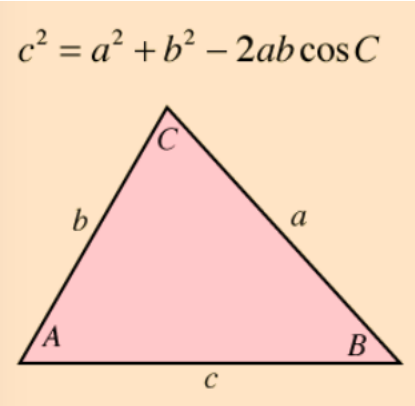
We can calculate similarly for $s + t = 5\pi/12$ or just switch sine and cosine from $\pi/12$.

Chapter 26

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



$$c^2 = a^2 + b^2 - 2ab\cos C$$

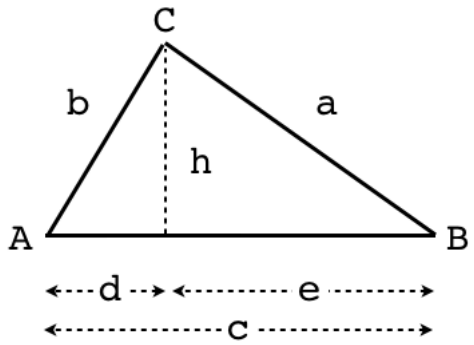
Lockhart calls this the "generalized" Pythagorean theorem. We can view the term $-2ab\cos C$ as a correction term which disappears in the

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.



$$a^2 - e^2 = h^2 = b^2 - d^2$$

So

$$a^2 = e^2 + b^2 - d^2$$

Since $d = c - e$ and thus $d^2 = c^2 - 2ce + e^2$:

$$a^2 = e^2 + b^2 - (c^2 - 2ce + e^2)$$

$$= b^2 - c^2 + 2ce$$

but $e = a \cos B$ so

$$a^2 = b^2 - c^2 + 2ac \cos B$$

rearrange to give a more familiar form (this is the law of cosines)

$$b^2 = a^2 + c^2 - 2ac\cos B$$

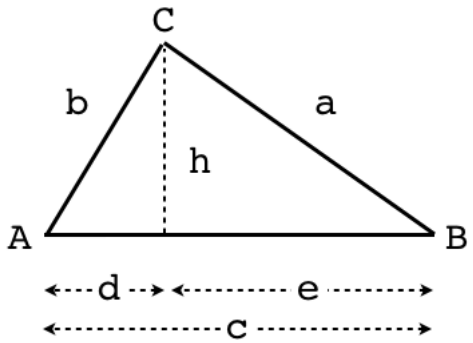
Any side of a triangle can be expressed in terms of the other two and the cosine of the angle between them. Thus, for example

$$c^2 = a^2 + b^2 - 2ab\cos C$$

$$a^2 = b^2 + c^2 - 2bc\cos A$$

Law of sines

I'll just mention that there is another law called the law of sines. In contrast to the law of cosines, it is fairly trivial.



$$\frac{h}{b} = \sin A \qquad \frac{h}{a} = \sin B$$

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 A and C or B and C. Therefore

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}$$

Part VII

Two basic operations in calculus

Chapter 27

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 will be watching your odometer.

Velocity times time = distance. We can think of speed and velocity as

the same for now. Distance divided by time is velocity.

Velocity is the *rate of change* of distance with time, it has units of distance divided by time (say, miles per hour).

In calculus we say that the velocity is the **derivative** of the distance with respect to time, and the 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.

time-dependence

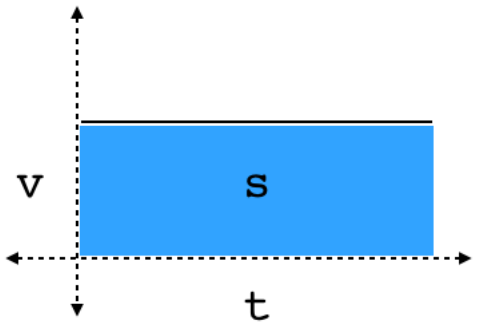
Distance equals velocity times time.

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

$$s = vt$$

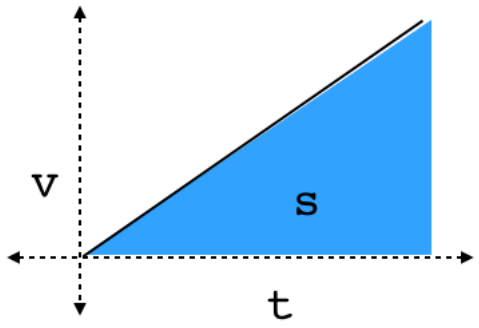
However, for most interesting problems velocity is not constant.

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?

If a is not zero then v changes 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:

$$s = v_{\text{avg}}\,t = \frac{1}{2}\,v \cdot t$$

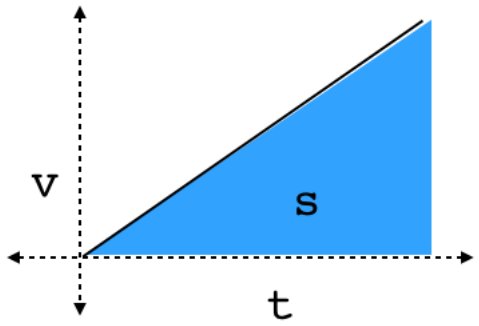
and since $v = at$

$$s = \frac{1}{2}at^2$$

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:

$$s = \frac{1}{2}gt^2$$

where g is the acceleration due to gravity.

Notice that this is the same equation as we had earlier. The reason is that g is approximately constant near the surface of the earth. g is about 10 in units of $\mathrm{m/s^2}$. 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

$$s = \frac{1}{2}at^2 + v_0t + s_0$$

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.

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.

However, in the first case, $y = c$, y does not actually depend on x at all. The result (dy, the change in y for a change in x, dx) is zero.

$$y = c, \qquad dy = 0 \cdot dx$$

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:

$$\frac{dy}{dx} = 0$$

The plot is a horizontal line with slope 0.

In the second case, y is a linear function of x, the change in y, 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 $(x_1, y_1), (x_2, y_2)$ we choose for the calculation. The following is true for *any* two pairs (x, y) on the line:

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

Above we had the example where $v = at$ with constant a. Then $dv/dt = a$.

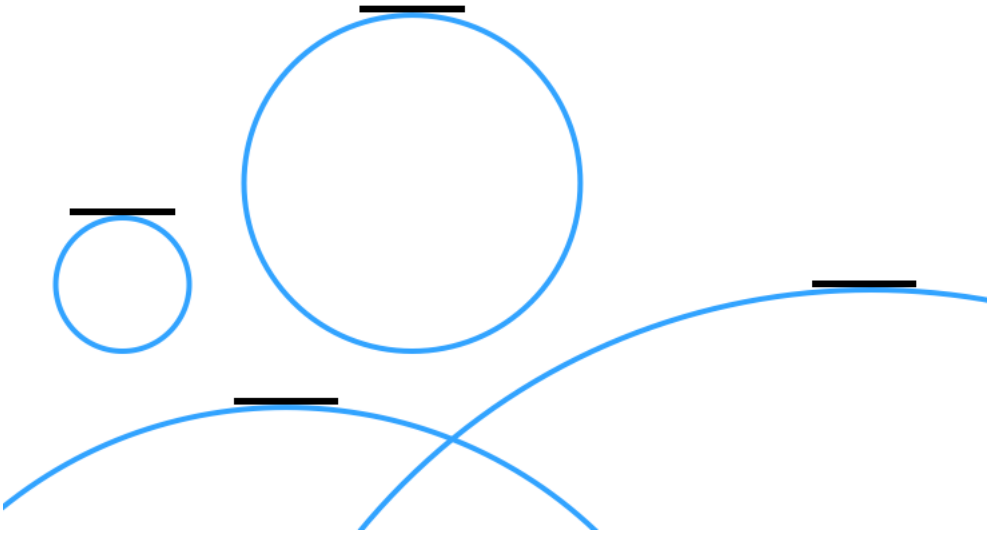
The third case is different.

$$y = cx^2$$

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.

It seems impossible to compute the slope of this curve in the standard way, by picking a second point near (x, y) and then calculating $\Delta y/\Delta x$, because the slope changes as we go out along the curve.

The key insight is that if x_1 is sufficiently close to x_2 the slope is constant. It's like saying that the earth is flat *locally*. If you detect any curvature, just zoom in a bit. In the figure



the line has constant length, but the distance to the circle from the end of the line decreases as we increase the size of the circle.

In calculus, we keep the curve the same size and decrease the length of the line, and then magnify the whole picture, until we get something like the figure above.

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 other words, 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.

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 $r - b$

```
r = 21120000
a =       50
b = 21119999.9999408
```

That is 0.00006 feet over the length of a 50 foot house, much much less than 1/16 of an inch. Roughly 6 parts in 10000 compared to 1 part in 192.

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 28

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

$$A = \int dA = \int \frac{1}{2} R\ R\ d\theta$$

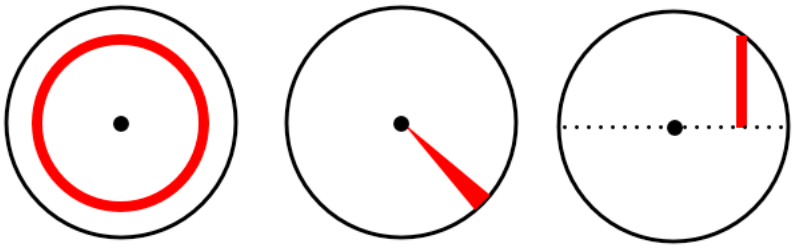
again we see that constants can come outside the integral

$$= \frac{1}{2} R^2 \int_{\theta=0}^{\theta=2\pi} d\theta$$

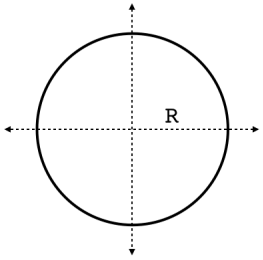
$$= \frac{1}{2} R^2 \theta \Big|_{\theta=0}^{\theta=2\pi}$$

$$= \pi R^2$$

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 ($r = 0 \rightarrow R$). 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

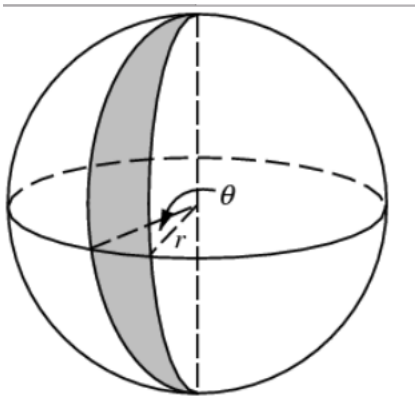
$$dV = 4\pi r^2 \, dr$$

$$V = \int dV = \int_0^R 4\pi r^2 \, dr$$

$$= 4\pi \left. \frac{1}{3} r^3 \right|_0^R = \frac{4}{3}\pi R^3$$

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

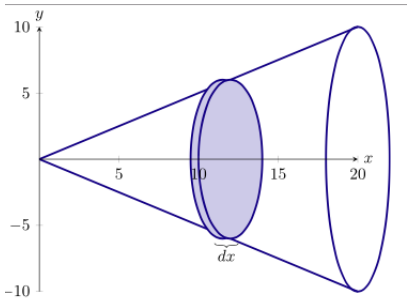
We should point out that this connection between volume and surface area is not true for *every* solid.

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**).