

Analytic Geometry and Trigonometry

Tom Elliott

December 25, 2020

Contents

I	Introduction	4
1	Introduction	5
II	Analytic geometry	7
2	Lines and slopes	8
3	Circles again	15
4	Parabola	31
5	Parabola slope	41
6	Ellipse	57
7	Hyperbola	71
III	Trigonometry	76
8	Six functions	77
9	Sum of angles	85
10	Double angle	94
11	Law of cosines	101
12	Heron and Brahmagupta	107

IV	Two basic operations in calculus	119
13	Simple slopes	120
14	Difference quotient	131
15	Easy pieces	141
16	Limit concept	152
V	Vectors	160
17	Vector dot product	161
18	Vector cross product	175
19	Point and plane	180
20	Headlight problem	185
21	Rotation	190
VI	Theta and r	199
22	Polar coordinates	200
23	Polar conics	204
24	Polar hyperbola	213
VII	Advanced problems	218
25	Langley	219
26	Circular arch	222
27	Hatbox theorem	227
28	Spherical cap	232

29	Pappus	236
VIII Addendum		242
30	More on sum of angles	243
31	Pythagorean triples	246
32	Archimedes and pi	248
33	Value of pi	263
34	Value of pi revisited	277
35	Square root of 3	293
36	Archimedes and quadrature	299
37	References	306

Part I

Introduction

Chapter 1

Introduction

This book has been modified from a previous project, where it comprised some chapters of a book entitled *Best of Calculus*. That one is here:

https://github.com/telliott99/calculus_book

I decided to make this material separate because the overall size of the big book made it hard to focus on the Geometry topics. These chapters have been the recipient of quite a bit of expansion and polishing.

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



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

Elements.

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

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

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

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

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

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

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

<https://github.com/telliott99/geometry>

Part II

Analytic geometry

Chapter 2

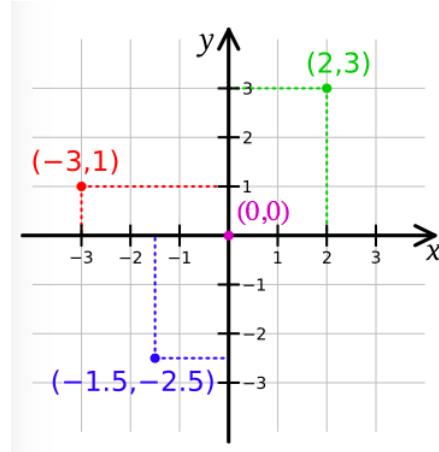
Lines and slopes

It is difficult today to put ourselves in the place of those who tried to reason about mathematics through the ages.

The Greeks lacked algebra, and although the Romans worked with numbers they did not have decimal notation. The concept of 0 came much later (from India), and even in the Middle Ages there was as yet no such thing as the equals sign $=$, which dates from 1557.

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

The invention of analytic geometry is often ascribed solely to Descartes, but Fermat also had his own version. There are two fundamental ideas.



The first is to orient two number lines on a piece of paper, at right angles, and then consider pairs of numbers (x, y) in the 2D plane. Such pairs or tuples are called the *coordinates* of points in the plane.

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 wearis 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 $y = x^2$. It turns out that this equation generates all the points of a parabola defined by classical criteria (namely, the distances between points on the curve and a point called the focus and a line called the directrix).

As a simple example, pick a few values of x and calculate the corresponding values of y . For example: $(0, 0), (\pm 1, 1), (\pm 2, 4), \dots$. Plot these points, and then finally, sketch the graph of the curve, without actually trying to plot *all* of the individual points (of which there is an infinite number). We make the assumption here that the function being plotted is continuous, so that the sketch of a curve between two points that are close enough together will be fairly smooth and if the x -values are close to the plotted x , the corresponding y -values will not be not too different from the plotted y .

point

A point is simply an ordered pair (x, y) such as $(1, 3)$. Often points have integer components, but they don't have to be.

distance formula

The x - and y -axes are perpendicular to one another (a fancy word for that is *orthogonal*).

Suppose we pick two particular points (s, t) and (u, v) , plot them on a graph, and then draw the line that connects them. Recall Euclid's first two postulates:

- A straight line segment can be drawn joining any two points.
- Any straight line segment can be extended indefinitely in a straight line.

The distance between the two points is given by the Pythagorean formula, where Δx is the change in x and Δy is the change in y :

$$d = \sqrt{\Delta x^2 + \Delta y^2}$$

It is often easier to use the squared distance and avoid the square root:

$$\begin{aligned} d^2 &= \Delta x^2 + \Delta y^2 \\ &= (s - u)^2 + (t - v)^2 \end{aligned}$$

Switching the order of (s, t) and (u, v) doesn't change the result.

That's because

$$(s - u)^2 = s^2 - 2su + u^2 = (u - s)^2$$

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

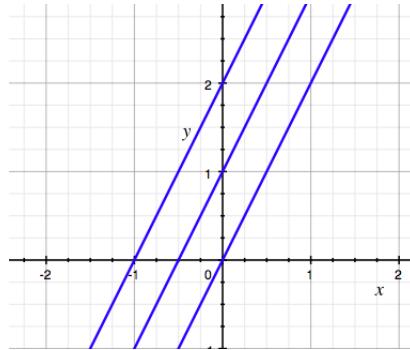
$$y = mx + y_0$$

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.

Another way to do this is to add a multiple of one equation to the other, such that one of the variables disappears. Here there are two possibilities:

$$y = 2x - 1$$

$$2y = -2x + 16$$

Addition gives $3y = 15, y = 5$. Alternatively,

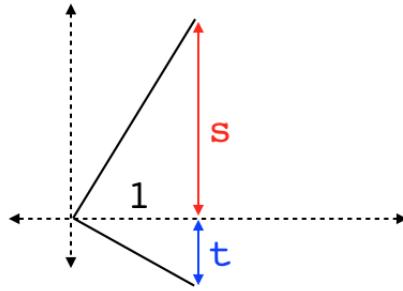
$$y = 2x - 1$$

$$-y = x - 8$$

Addition gives $3x - 9 = 0, x = 3$.

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 .



Proof.

Draw two lines going through the origin, forming a right angle there. The first has slope s , so it goes through the point $(1, s)$, the second has slope $-t$ and goes through $(1, -t)$.

As a corollary of the Pythagorean theorem, we found that the product of the two pieces of the base is equal to the altitude squared.

Here:

$$st = 1^2 = 1$$

These are the lengths, i.e. the absolute values of the slopes.

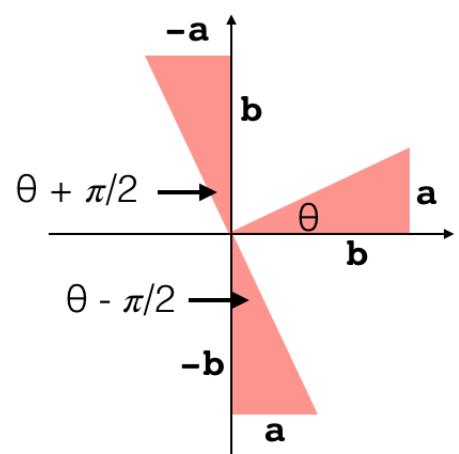
$$|s| \cdot |t| = 1$$

But clearly the sign of t is negative. So we arrive at

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

$$m_1 \cdot m_2 = -1$$

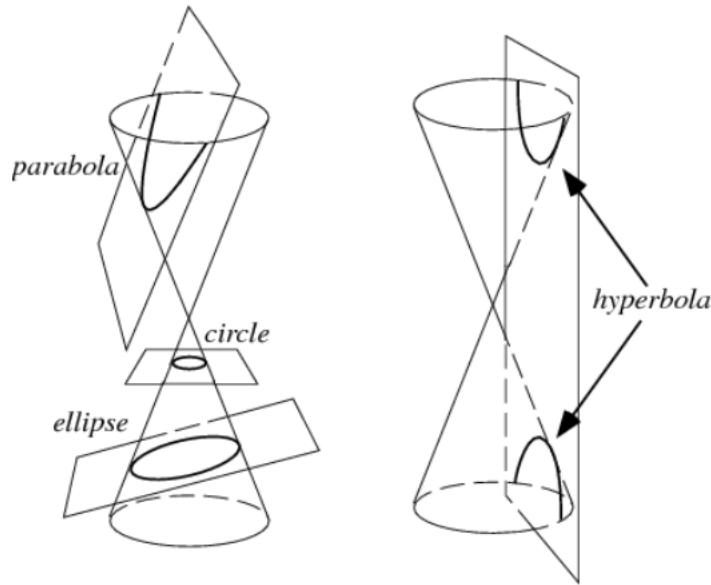
We'll see a natural easy proof of this once we look at trigonometry. Here is a hint:



Chapter 3

Circles again

We now consider what are called quadratic forms, as distinguished from linear equations (i.e., for lines). Quadratics contain one or two squared terms (or a term that mixes x and y).



One of the simplest examples is the equation for a unit circle centered at the origin:

$$x^2 + y^2 = 1$$

Pythagoras tells us that for a point (x, y) , the square of the distance from the origin

is $x^2 + y^2$. This equation describes all the points whose distance from the origin is equal to $\sqrt{1} = 1$. But all the points equi-distant to a point form a circle. We generalize

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

It is clear that when $y = 0$, $x = \pm r$, and when $x = 0$, $y = \pm r$. r is the radius of the circle.

shifted circle

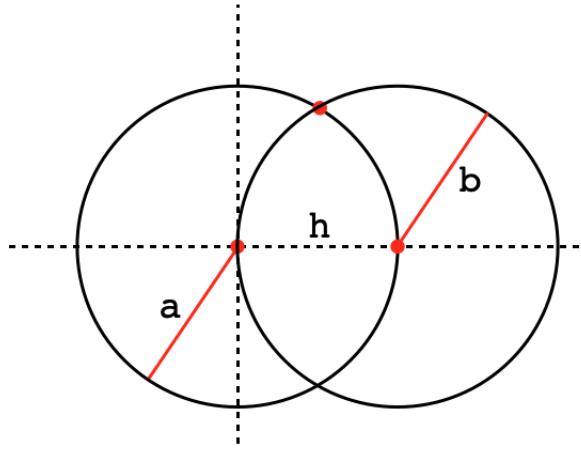
If the center is at $(1, 0)$, what this amounts to is adding 1 to the x value of every point. If we solve for x

$$x = \sqrt{1 - y^2}$$

and then add 1

$$\begin{aligned} x &= \sqrt{1 - y^2} + 1 \\ (x - 1)^2 &= 1 - y^2 \\ (x - 1)^2 + y^2 &= 1 \end{aligned}$$

So even though the displacement is positive, say a value h , the x term becomes $(x - h)$, which is counter-intuitive.



Here are two circles, one at the origin with radius a , and one at $(h, 0)$ with radius equal to b . The equation of the displaced one is:

$$(x' - 1)^2 + y'^2 = b^2$$

$$x'^2 - 2x' + 1 + y'^2 = b^2$$

and the other is of course:

$$x^2 + y^2 = a^2$$

To find the points where the two circles intersect, we must have $x = x'$ and $y = y'$:

$$x^2 - 2x + 1 + y^2 = b^2$$

$$x^2 + y^2 = a^2$$

Subtract to obtain:

$$-2x + 1 = b^2 - a^2$$

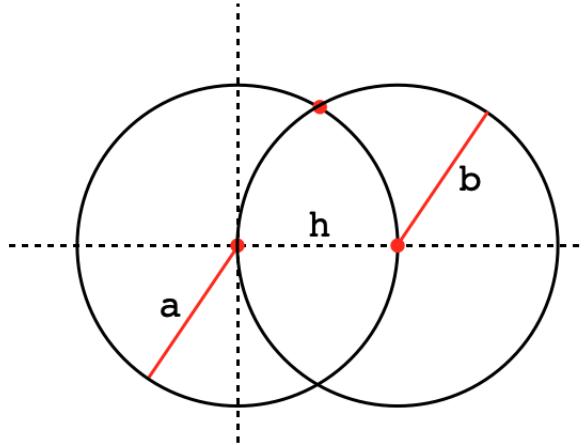
$$2x - 1 = a^2 - b^2$$

If $a = b = 1$ then

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

and

$$y = \pm\sqrt{1 - x^2} = \pm\frac{\sqrt{3}}{2}$$



Notice that the distance from the origin to the point of intersection is

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

$$= 1$$

But that's just equal to h .

We have found the vertex of an equilateral triangle. Since you remember **this**, so that's no surprise.

For another example, let $a = 2$, and everything else stay the same. Then

$$2x - 1 = 2^2 - 1^2$$

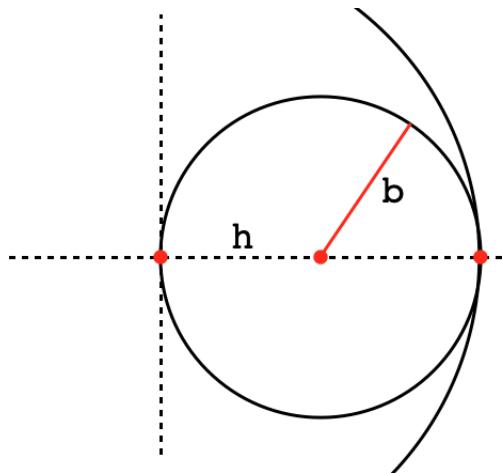
$$2x = 4$$

$$x = 2$$

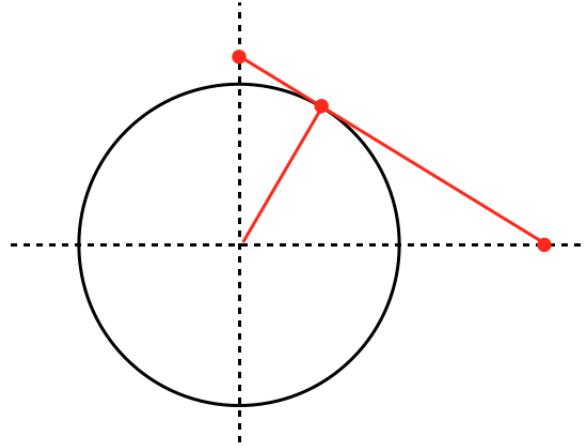
So

$$\begin{aligned}y &= \pm\sqrt{r^2 - x^2} \\&= \pm\sqrt{2^2 - 2^2} = 0\end{aligned}$$

And if a were larger than twice h , we get to the last step and find the square root of a negative number, because the circles don't intersect.



tangent to the circle



Let's consider an arbitrary point on a circle (x, y) . The radius squared is equal to $x^2 + y^2$.

We draw the radius to the point (x, y) and also the tangent at the same point, and recall an important fact from plane geometry: the tangent and the radius meet at a right angle.

Proof.

Of course, we could just appeal to symmetry.

The tangent is defined as the line that touches only a single point on the circle. That point is closer to the origin of the circle than any of the other points on the tangent line, since none of them touch the circle and so their distance to the origin is greater than the distance to the tangent point.

We proved in the chapter on right triangles that the distance from a point to a line is least at the point where the new segment makes a right angle with the line.

□

The slope of the radius to that point is just y/x . From our work with lines and slopes we know that perpendicular (orthogonal) lines have slopes whose product is -1 . So the slope of the tangent line is $-x/y$.

We use the point-slope equation for the two points (x, y) , and $(x_0, 0)$ to write:

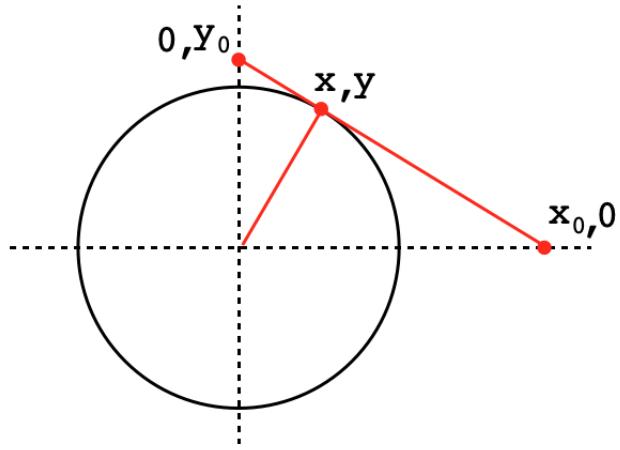
$$-\frac{x}{y} = \frac{0 - y}{x_0 - x}$$

$$x(x_0 - x) = y^2$$

$$xx_0 = r^2$$

For a unit circle, $x_0 = 1/x$

This makes sense, when $x = r$, then $x_0 = 1/r$ and as $x \rightarrow 0$, $x_0 \rightarrow \infty$.



Similarly, the line between $(0, y_0)$ and (x, y) gives:

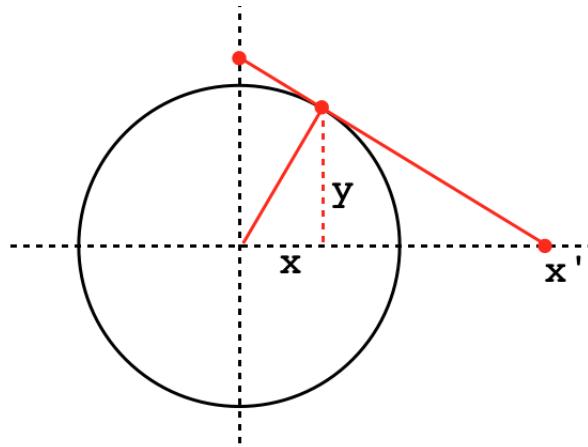
$$-\frac{x}{y} = \frac{y - y_0}{x - 0}$$

$$y(y - y_0) = -x^2$$

$$r^2 = yy_0$$

For a unit circle, $y_0 = 1/y$.

You can also do this using the Pythagorean theorem.



The radius squared is $x^2 + y^2$. The length of the segment from (x, y) to $(x', 0)$ is

$$(x' - x)^2 + y^2$$

And the hypotenuse on the bottom is x'^2 . So

$$\begin{aligned} x'^2 &= (x' - x)^2 + y^2 + x^2 + y^2 \\ &= x'^2 - 2x'x + x^2 + y^2 + x^2 + y^2 \\ &= x'^2 - 2x'x + 2r^2 \end{aligned}$$

Thus

$$2r^2 = 2x'x$$

$$rr = x'x$$

displaced circle

Let's go back to the displaced circle (not centered at the origin). Generally

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

where the origin of the circle is at (h, k) .

Multiplying out:

$$\begin{aligned} x^2 - 2hx + h^2 + y^2 - 2ky + k^2 &= r^2 \\ x^2 + y^2 - 2hx - 2ky + (h^2 + k^2 - r^2) \end{aligned}$$

Comparing to the most general form for a quadratic

$$Ax^2 + By^2 + Cxy + Dx + Ey + F = 0$$

We see that

$$A = 1, \quad B = 1, \quad C = 0$$

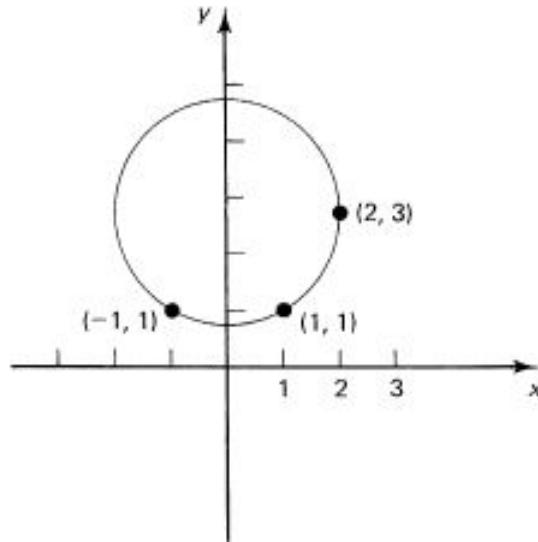
and in fact, this is true for all circles. (If $A = B \neq 1$, just divide all the terms by A).

Moreover

$$D = -2h, \quad E = -2k, \quad F = h^2 + k^2 - r^2$$

This equation can help us solve the following problem from Hamming: find the equation of the circle that passes through the following three points:

$$(-1, 1), (1, 1), (2, 3)$$



We write

$$x^2 + y^2 + Dx + Ey + F = 0$$

From the values of x and y at each of the three points we get

$$1 + 1 - D + E + F = 0$$

$$1 + 1 + D + E + F = 0$$

$$4 + 9 + 2D + 3E + F = 0$$

Three equations in three unknowns. We can do that.

Adding the first two equations together:

$$4 + 2(E + F) = 0$$

so $E + F = -2$.

Subtracting the first two equations (or substituting the result for $E + F$) tells us that $D = 0$.

Adding (-3) times the second equation to the third gives:

$$1 + 6 - D - 2F = 0$$

$$7 - 2F = 0$$

$F = 7/2$, and since $E + F = -2$, $E = -11/2$.

So the solution is

$$x^2 + y^2 - \frac{11}{2}y + \frac{7}{2} = 0$$

You can check that it works for all three points:

$$(-1, 1), (1, 1), (2, 3)$$

The first two are easy, while the third gives

$$4 + 9 - \frac{11}{2}3 + \frac{7}{2} = 0$$

$$8 + 18 - 33 + 7 = 0$$

which looks correct.

completing the square

We can improve this by completing the square. We see that

$$y^2 - \frac{11}{2}y + \left(\frac{11}{4}\right)^2 = \left(y - \frac{11}{4}\right)^2$$

We must add that back to the right-hand side of the original to obtain:

$$x^2 + \left(y - \frac{11}{4}\right)^2 = \left(\frac{11}{4}\right)^2 - \frac{7}{2}$$

The center is at $(0, 11/4)$. The radius doesn't come out cleanly but r^2 is

$$\frac{121}{16} - \frac{56}{16} = \frac{65}{16}$$

so r is slightly more than 2.

Or recall that we had:

$$D = -2h, \quad E = -2k, \quad F = h^2 + k^2 - r^2$$

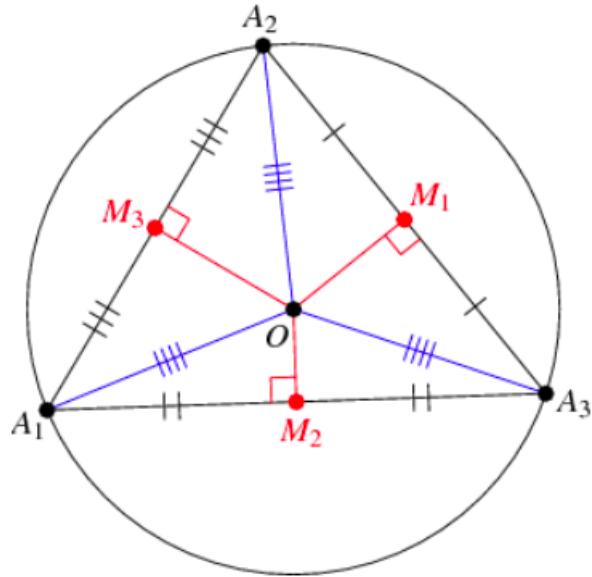
From this, we have that $h = 0$ and $k = -E/2 = 11/4$, and the radius is more complicated, as we said.

plane geometry

We can check our work by solving the problem using a technique from plane geometry. Again, we want the circle passing through three points:

$$(-1, 1), (1, 1), (2, 3)$$

Take two of the points to be placed on a circle and construct the line segment joining them (a chord of the circle). Find the midpoint of the chord and erect a perpendicular bisector through the midpoint. Now, every point lying on the bisector is equidistant from the two starting points. Proof: draw the two triangles including that point, the two starting points and the midpoint of the bisector. The two triangles are congruent. Here is the general picture.

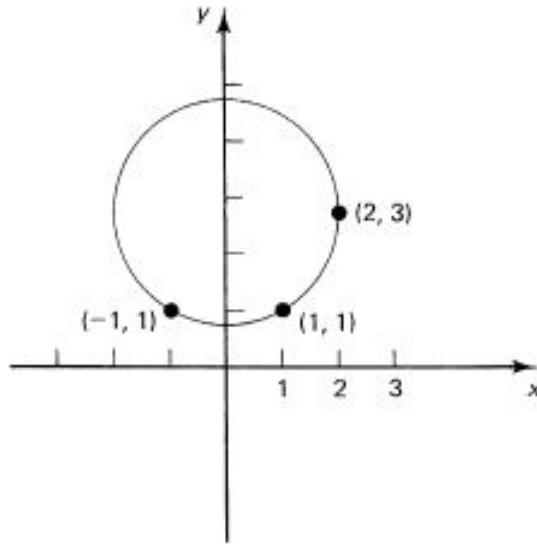


It's a bit trickier to prove that *every* point that is equidistant from the two points lies on the bisector. We assume that.

Since every point that is equidistant from the two points lies on the bisector, the radius of the circle lies on the bisector.

Then, erect a perpendicular bisector of a chord joining another pair chosen from the three points. This new bisector and the first one meet at the center of the circle.

In our case two points $(-1, 1), (1, 1)$ are symmetric about the y -axis. Therefore it is clear that the perpendicular bisector for these two points is the y -axis.



For the second bisector, form the vector between $(1, 1)$ and $(2, 3)$ as $\mathbf{v} = \langle 1, 2 \rangle$. The midpoint is at $(1, 1) + \mathbf{v}/2 = (3/2, 2)$.

The slope of the bisector is the negative inverse of the slope for the chord which is $-1/2$ so the equation of the bisector is

$$y - y_0 = -\frac{1}{2}(x - x_0)$$

Plugging in the point that we know, we obtain

$$y - 2 = -\frac{1}{2}(x - 3/2)$$

We want to solve for y when $x = 0$, crossing the first bisector, the y -axis

$$\begin{aligned} y - 2 &= -\frac{1}{2}(-3/2) \\ y &= \frac{11}{4} \end{aligned}$$

So the center is at $(0, 11/4)$, which matches what we had before. We compute the distance to one of the points $(1, 1)$ as

$$d = \sqrt{1^2 + (11/4 - 1)^2} = \sqrt{1 + 49/16}$$

which also matches our previous result.

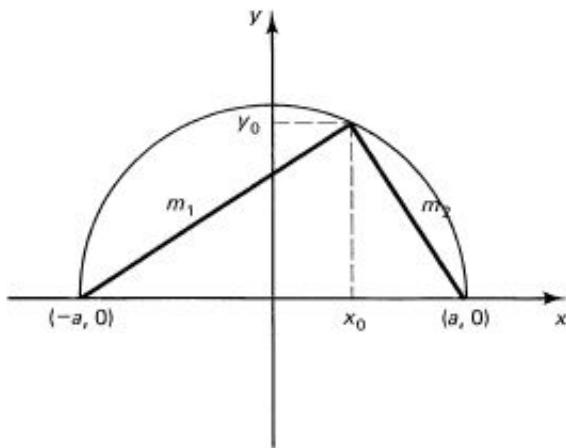


Figure 6.2-3 Angle in a semicircle

Here is another problem from Hamming. We need to prove that the angle above is a right angle. We know this is true from geometry. But now we wish to practice our analytic geometry.

Suppose the equation of the circle is

$$x^2 + y^2 = a^2$$

The point on the circle is (x_0, y_0) .

Our first solution uses slopes and points. The line from $(-a, 0)$ to (x_0, y_0) has slope

$$m_1 = \frac{y_0}{x_0 + a}$$

The line from $(a, 0)$ to (x_0, y_0) has slope

$$m_2 = \frac{y_0}{a - x_0}$$

Two lines meet at a right angle if the product of their slopes is equal to -1 .

$$m_1 m_2 = \frac{y_0}{x_0 + a} \cdot \frac{y_0}{a - x_0}$$

$$= \frac{y_0^2}{a^2 - x_0^2} = \frac{y_0^2}{x_0^2 + y_0^2 - x_0^2} = -1$$

This was not pretty, it's just good exercise.

And here is a proof using vectors and the dot product. Consider the semicircle centered on the origin with radius a , so the ends of the diameter are at $(x = \pm a, 0)$.

Form the vectors from those ends to an arbitrary point (x, y) on the perimeter:

$$\mathbf{u} = \langle x + a, y \rangle$$

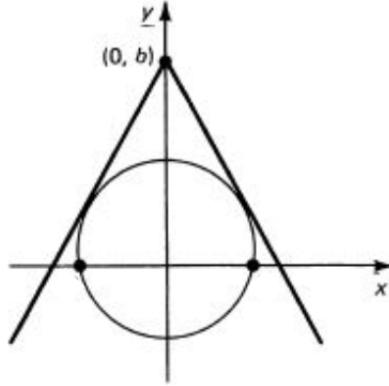
$$\mathbf{v} = \langle x - a, y \rangle$$

Notice that

$$\begin{aligned}\mathbf{u} \cdot \mathbf{v} &= (x + a)(x - a) + y^2 \\ &= x^2 - a^2 + y^2 = 0\end{aligned}$$

because $x^2 + y^2 = a^2$ for any point on the circle.

As our last example, consider the problem of finding the equation of a line tangent to a circle that goes through some arbitrary point b .



We take the circle to have radius a and be centered at the origin. We take the point b to be on the y -axis. The equation of the line on the right side is

$$\frac{y - y_0}{x - x_0} = m = \frac{y - b}{x}$$

$$y = mx + b$$

(well, of course).

For the point or points where the line intersects the circle we also have

$$\begin{aligned}y &= \sqrt{a^2 - x^2} \\ \sqrt{a^2 - x^2} &= mx + b \\ a^2 - x^2 &= m^2 x^2 + 2bmx + b^2 \\ (m^2 + 1)x^2 + 2bmx + b^2 - a^2 &= 0\end{aligned}$$

From the quadratic equation:

$$x = \frac{-2bm \pm \sqrt{4b^2m^2 - 4(m^2 + 1)(b^2 - a^2)}}{2(m^2 + 1)}$$

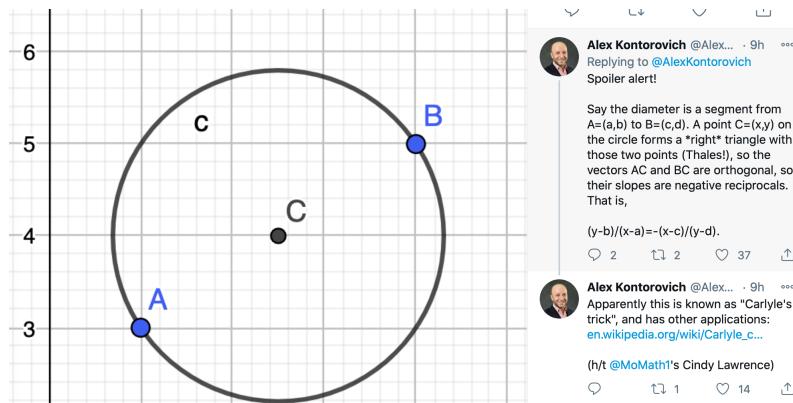
We are looking for the case where there is a single solution so the discriminant under the square root must be equal to zero:

$$\begin{aligned}4b^2m^2 &= 4(m^2 + 1)(b^2 - a^2) \\ m^2b^2 &= m^2b^2 - m^2a^2 + b^2 - a^2 \\ 0 &= -m^2a^2 + b^2 - a^2 \\ m &= \pm \frac{\sqrt{b^2 - a^2}}{a}\end{aligned}$$

This makes sense since if $a = b$ the single tangent should be horizontal with zero slope. Notice that if $a^2 > b^2$ there is no real solution. This corresponds to having b inside the circle.

one more

And then lastly, I found this problem on the web:



The challenge was, can you immediately write the equation of this circle? I said sure

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

$$(x - 2.5)^2 + (y - 4)^2 = 1.8^2$$

But notice we are given $A = (1, 3)$ and $B = 4, 5)$.

There is a reason! According Thales' theorem, every point (x, y) on the circle should have a vector to A and a vector to B , that when multiplied to give the dot product, you get zero.

Write:

$$\langle(x - 1), (y - 3)\rangle \cdot \langle(x - 4), (y - 5)\rangle = 0$$

$$x^2 - 5x + 4 + y^2 - 8y + 15 = 0$$

We need to complete *two* squares:

$$x^2 - 5x + 2.5^2 + y^2 - 8y + 4^2 + \dots$$

$2.5^2 = 6.25$ and $4^2 = 16$ and the sum is 22.25. Since we originally had 19 we now have -3.25 on the left and write

$$(x - 2.5)^2 + (y - 4)^2 = 3.25$$

$\sqrt{3.25} \neq 1.8$. What went wrong?

Calculate the distance between A and B as $\sqrt{3^2 + 2^2} = \sqrt{13} = 3.60555\dots$. Half of that is the radius, which is 1.802775....

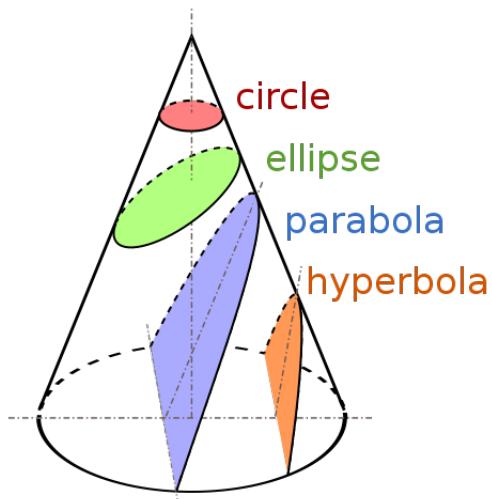
And that matches $\sqrt{3.25} = 1.802775\dots$

The problem was that although the upper edge of the circle looked to be at $4 + 1.8 = 5.8$, that is not exactly correct. The statement was that A and B have integer values for (x, y) .

Chapter 4

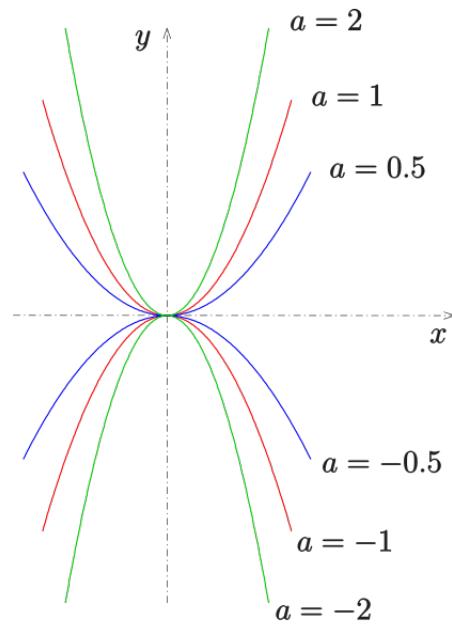
Parabola

The parabola is one of a larger class of geometric figures called the conic sections.

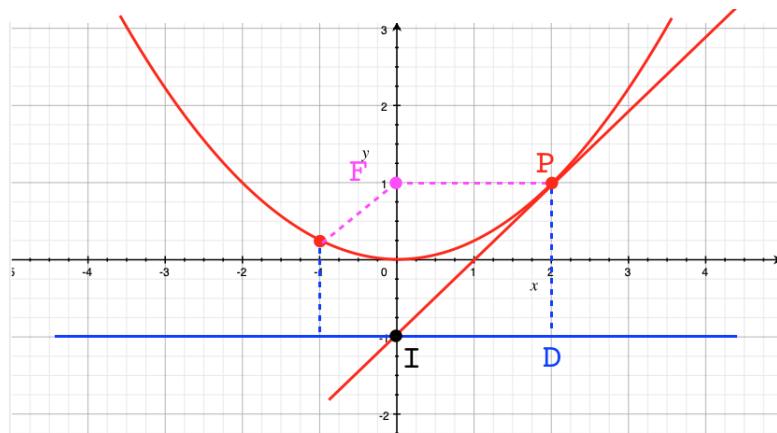


It is pretty complicated to look at parabolas in the way that the Greeks did, so we start by using a formula from analytic geometry, which allows us to draw the curve.

The equation of a general parabola is simply $y = ax^2$, where a is called the shape factor of the parabola. Here are parabolas with different values of a . The larger a is, the faster the curve rises.



The one in the figure below is a little flatter than we usually draw, we'll see the reason for that choice in a minute. The vertex is at $(0, 0)$, and the curve is symmetric across the y -axis: $f(x) = f(-x)$.

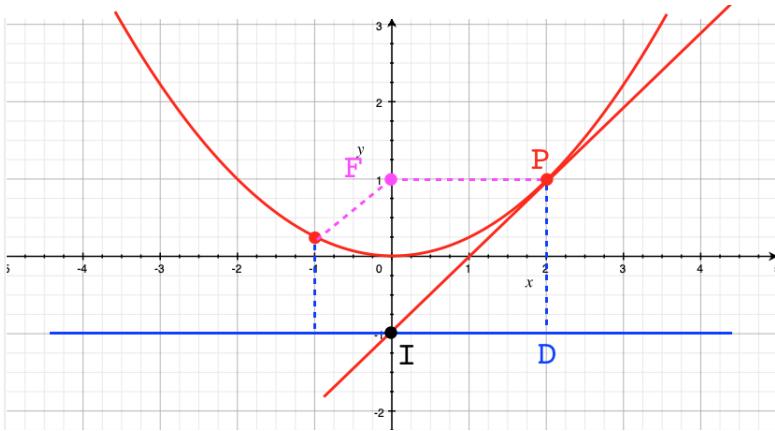


At $x = 2$ we have $y = 1$, so $a = 1/4$. And at $x = 1$ we have $y = 1/4$, which is consistent with that.

focus and directrix

We need one more idea to start our geometric look at the parabola.

The geometric definition is this. 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 (F).



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 blue and its equation is $y = -p$.

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

It is another fact that we will establish later that the distance p is related to a by the equation:

$$4ap = 1$$

which explains our choice of a . We want the parabola to be flat enough to see F and the line $y = -p$ clearly.

If we consider the point $P = (2, 1)$ we can compute the distance to the focus as simply $\Delta x = 2$ and to the directrix as $\Delta y = 1 + 1 = 2$.

slope of the tangent

One last fact we will justify later: at any point (x, y) on the parabola, the slope is $2ax$. Therefore, the slope of the tangent to the curve at $x = 2$ is

$$m = 2 \cdot 1/4 \cdot 2 = 1$$

By inspection of the graph we see that the tangent line goes through $I = (0, -1)$ which gives a point slope formula of $y = x - 1$. It is easy to verify that $(2, 1)$ and $(0, -1)$ are both on the line, and of course the slope is 1, as advertised.

Notice that the x -intercept, x_0 is

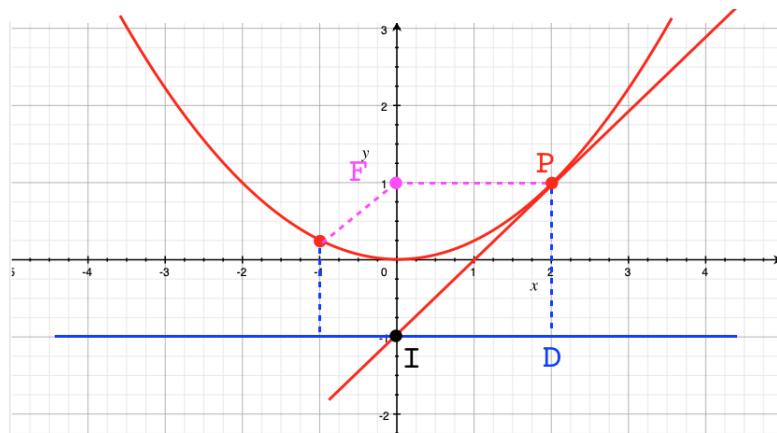
$$0 = x - 1, \quad x = 1$$

This is exactly halfway on the x -axis between P and I .

All of this depends on our assumption that the slope of the tangent to the curve at $(2, 1)$ has slope $2ax = 1$ for $y = 1/4 x^2$.

another point

From the equation of the curve we can get that $(x = -1, y = 1/4)$ is on the curve. The distance from the point to the directrix is just $5/4$.



The distance to the focus, squared, is:

$$d^2 = 1^2 + \left(\frac{3}{4}\right)^2 = \frac{25}{16}$$

so

$$d = \sqrt{\frac{25}{16}} = \frac{5}{4}$$

which checks.

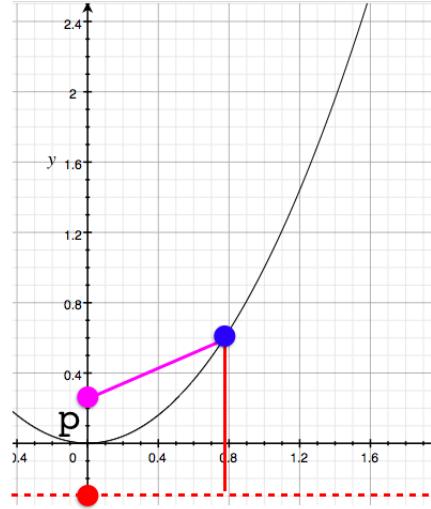
And this should not be a surprise. A look at the figure will show that in units of $1/4$, we have a 3-4-5 right triangle.

computing p

Pick an arbitrary point on a parabola (in blue), with coordinates (x, ax^2) .

The change in x going to the focus is just x , while going to the directrix it is zero.

To compute the change in y , find the distance of the point to the x -axis as ax^2 and then for the focus, subtract p , while for the directrix, add p .



So, the squared distance to the focus (magenta point) is

$$\Delta x^2 + \Delta y^2 = x^2 + (ax^2 - p)^2$$

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

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

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

At this point one can multiply out the squares on both sides. Notice, we have $(m-n)^2$ on the left and $(m+n)^2$ on the right. The terms of m^2 and n^2 will cancel leaving terms like $2mn$ of differing sign:

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

Divide by x^2

$$-2ap + 1 = 2ap$$

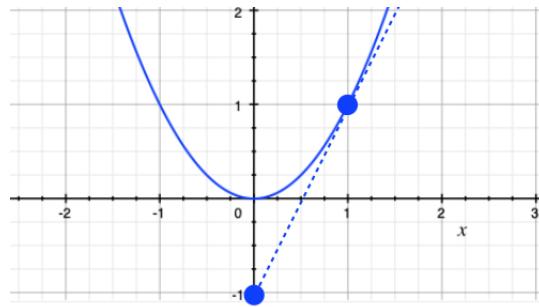
$$4ap = 1$$

$$ap = \frac{1}{4}$$

The shape factor a determines the distance of the focus and directrix from the vertex.

slope of the tangent

As we said, the slope of the tangent to $y = ax^2$ at any fixed point x is equal to $2ax$.



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 of the tangent line ($y' = 0$, $x' = x_0$) is:

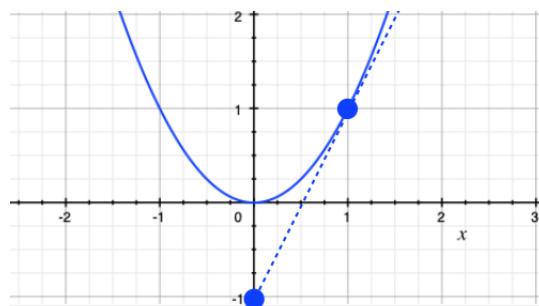
$$-ax^2 = 2axx_0 - 2ax^2$$

$$ax^2 = 2axx_0$$

$$x = 2x_0$$

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

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

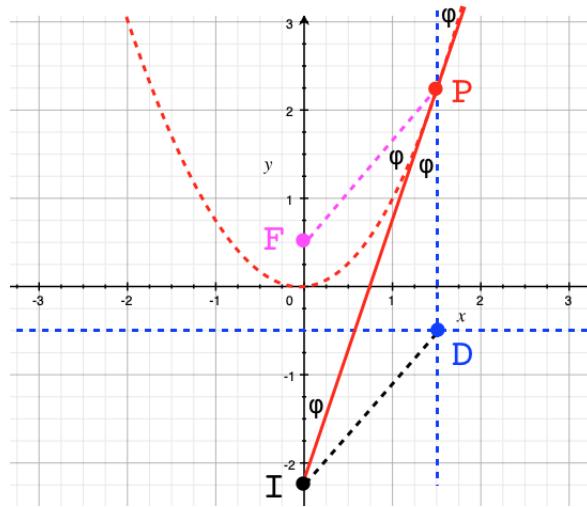


And what *that* means is that the y -intercept is symmetrical with the original point (as far below the x -axis as the point is above it). Here's the algebra:

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

$$y_0 = -ax^2$$

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 parallelogram. It looks like all four equal sides are equal. Let's see.

Proof.

The vertical distance from P to the x -axis is ax^2 , so $PD = ax^2 + p$. Similarly, the vertical distance from I up to the y -axis is also ax^2 , so $IF = ax^2 + p = PD$.

FP is the hypotenuse of a right triangle with sides x and $ax^2 - p$. ID is the hypotenuse of a right triangle with the same sides. Therefore $ID = FP$ and $IFPD$ is a parallelogram.

Finally, $FP = PD$, by the geometric definition of the parabola.

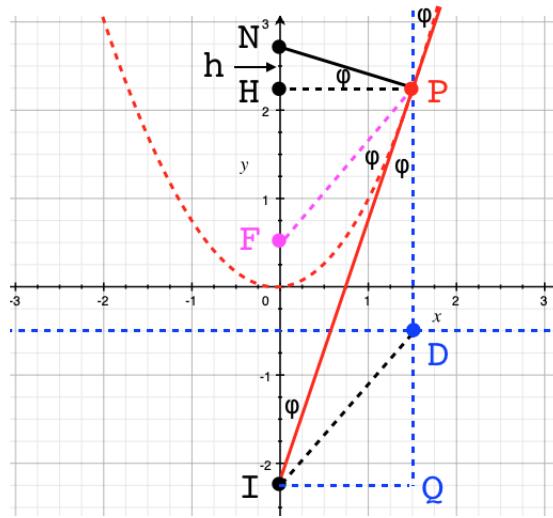
Therefore, $IFPD$ is a regular parallelogram.

□

As the long diagonal of a parallelogram, the tangent line makes equal angles with FP and PD .

If PD is extended vertically upward, the angle it makes with the tangent line (ϕ) is equal to the angle between FP and the tangent line. This means that all vertical light rays entering a parabola will reflect and then come together at the focus.

Let us add one more relationship to the figure: draw the line passing through P that is vertical to the tangent, called the *normal* to the tangent.



The normal intersects the y -axis at a point above P . We can calculate the additional vertical distance h by noting that the slope of this line is the negative inverse of $2ax$ and the distance along the x -axis is just x so

$$\Delta y = -\frac{2ax}{x} = -\frac{1}{2a}$$

where $h = |\Delta y|$. The y -intercept is $1/2a + ax^2$.

A more subtle way of doing this is to draw similar triangles on the figure. One in particular has base x and height h and hypotenuse equal to the length of the normal. If you do this, you should be able to show that

$$\frac{h}{x} = \frac{x}{2ax^2}$$

$$2ah = 1$$

which leads directly to the same answer.

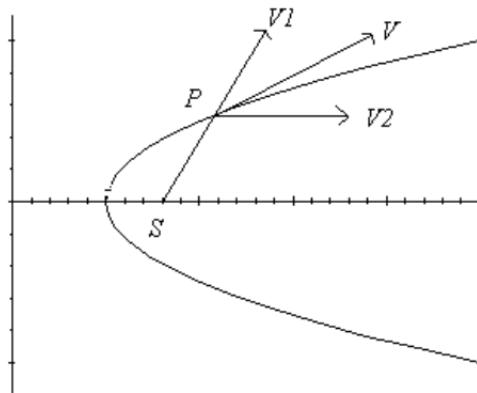
This means that the height of the intersection of the normal above the point (x, y) is independent of x and y . It's just $1/2a$. So as x gets larger and the parabola becomes more and more vertical, the distance in height gained by the normal becomes smaller and smaller as a fraction of x (or y).

Roberval's idea

Here is a brilliant idea of Roberval which gives the slope of the tangent to a parabola. It involves a tiny bit about vectors, which we will introduce later in the book.

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

Roberval views the curve as the track of an object, and asks what is the direction of movement at each position? (He draws his parabolas sideways, so we will too, for this once).



The idea is that the vector describing the motion, where a projectile tracing out the curve is headed at any individual moment, is the combination of two other vectors.

We know these two vectors! They are the invariants, namely that the distance to the focus and the distance to the directrix are equal at every point. Since the motion must be such as to preserve that invariant, the motion lies in the direction of the sum of the two vectors. These are

$$\begin{aligned} PF &= \langle x, ax^2 - p \rangle \\ PD &= \langle 0, ax^2 + p \rangle \end{aligned}$$

The sum is

$$\mathbf{v} = \langle x, 2ax^2 \rangle$$

A line along this vector has slope

$$\frac{\Delta y}{\Delta x} = \frac{2ax^2}{x} = 2ax$$

And since the tangent lies in the direction of motion, we're done. This general idea works for other curves, notably the circle and the ellipse.

Chapter 5

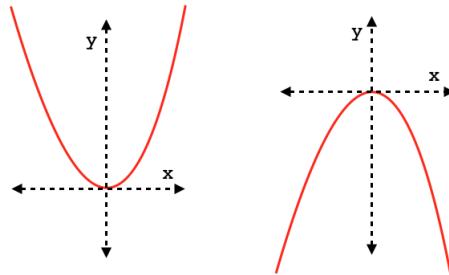
Parabola slope

direction of opening

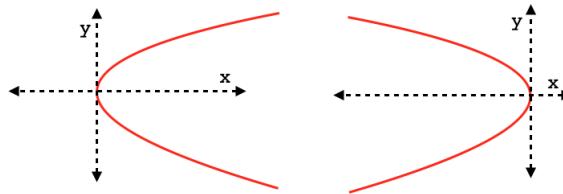
We continue looking at parabolas, the graphs of equations like $y = ax^2$, this time emphasizing the point of view of analytical geometry.

Any simple parabola (with no terms that mix x and y), has its axis of symmetry parallel to either the x - or y -axis.

Below we have $y = ax^2$ with either $a > 0$ (left panel) or $a < 0$ (right panel)



And then we have $x = ay^2$ with either $a > 0$ (left panel) or $a < 0$ (right panel)



In any case, we can just switch $-a$ and a , or x and y as needed, to obtain a figure like the first one, left panel. We focus on parabolas pointing up, with no loss of generality.

Those containing terms like Bxy , we leave aside as a complication to be returned to later.

vertex at different points

Start with a parabola having its vertex at the origin. The equation will be simply $y = ax^2$.

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

We can see that the vertex is at $(0, 0)$, because (i) $(0, 0)$ is on the curve and (ii) $y \geq 0$, so 0 is the smallest y .

It turns out that any parabola with its vertex at a point other than the origin, can be described by the formula

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

Moving the graph amounts to changing the values of x and y for every point on the curve.

For example, to move the vertex up by two units to $(0, 2)$, add 2 to the value of ax^2 for every x . The result is

$$y = ax^2 + k$$

where k is the amount of vertical shift. This can be rearranged to

$$(y - k) = ax^2$$

The vertex is at $y = 2$ (positive), but the formula says to subtract k from y , which is a bit counter-intuitive.

Changes in x are taken into account *before* squaring. For a parabola whose vertex is on the x -axis, the formula becomes

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

If you try this for a vertex at $(1, 0)$, and plot the values, you will see that this is correct. For example, x values symmetric on each side of the vertex yield the same y , in the proportion $a(\Delta x)^2$, where $\Delta x = x - h$.

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

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

If we work with this a bit, multiplying out:

$$\begin{aligned} y - k &= a(x^2 - 2xh + h^2) \\ y &= ax^2 - 2ahx + ah^2 + k \end{aligned}$$

In this form the cofactors are usually simplified as

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

which we are used to seeing from algebra.

Comparing the two, we see that the cofactors of the x term must be equal:

$$\begin{aligned} -2ahx &= bx \\ h &= -\frac{b}{2a} \end{aligned}$$

and the constant terms must be equal as well

$$\begin{aligned} c &= ah^2 + k \\ k &= c - ah^2 \\ &= c - \frac{b^2}{4a} \end{aligned}$$

We can check this as follows. The first equation is commonly used to find the vertex for a given parabola. The x -value of the vertex is $h = -b/2a$.

Then the y -coordinate (k) can obtained by plugging into the given equation:

$$\begin{aligned} k &= a\left(-\frac{b}{2a}\right)^2 + b\left(-\frac{b}{2a}\right) + c \\ k - c &= \frac{b^2}{4a} - \frac{b^2}{2a} \\ &= -\frac{b^2}{4a} \end{aligned}$$

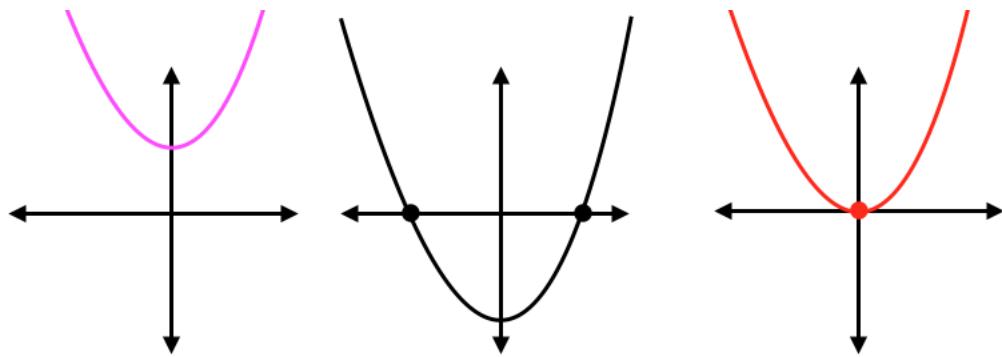
which matches what we had above.

roots

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

By trying different possibilities it becomes clear that it is possible to have 0, 1 or 2 roots.

In the figure below, the black curve has two roots, the red curve has one. The latter's equation is $y = x^2$, the former is $y = x^2 - 1$ and then we can see that $x^2 = 1$ has two real solutions $x = \pm 1$.



On the left, the magenta curve does not cross the y -axis. Its equation is $y = x^2 + 1$, and there are no (real) solutions, no 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 $a(x - s)^2 = 0$. A common example of that is a parabola with its vertex at the origin, so $s = 0$ and $y = ax^2$ (right panel, above).

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

completing the square

Suppose we have

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

When $y = 0$:

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

First, multiply through by $1/a$ and place the constant term on the right-hand side:

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

We want to write the left-hand side

$$x^2 + \frac{b}{a}x$$

as a perfect square. Something like

$$(x + p)^2 = x^2 + 2xp + p^2$$

What should p be?

If we compare the cofactor of the x term, in our problem we have b/a and in the example we have $2p$. So $2p = b/a$, $p = b/2a$ and then $(x + p)$ is like $(x + b/2a)$ so finally $(x + p)^2$ is like

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

The key insight is that on the original left-hand side (three equations back)

$$x^2 + \frac{b}{a}x$$

the third term is missing, but *we can fix it*. To maintain equality, simply add the same thing on both sides:

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

So now the left-hand side is a perfect square:

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

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

Now, we just do a bit of rearranging.

Multiplying top and bottom of the first term under the square root gives a common factor of $4a^2$:

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

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*:

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

This formula always works to find the roots of an equation, if they exist. The quantity under the square root is called the discriminant

$$D = b^2 - 4ac$$

If $D < 0$ then \sqrt{D} does not exist in the real numbers and there is no x such that $y = 0$. That corresponds to the case where the parabola does not cross the x -axis.

If $D = 0$ then there is a single root, and the graph just touches the x -axis.

check the answer

We assert that when x takes on those two values, it is a solution for

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

Take the positive root:

$$x = \frac{1}{2a} [-b + \sqrt{b^2 - 4ac}]$$

Compute ax^2

$$ax^2 = \frac{1}{4a} [b^2 - 2b\sqrt{b^2 - 4ac} + b^2 - 4ac]$$

$$= \frac{1}{4a} [2b^2 - 2b\sqrt{b^2 - 4ac} - 4ac]$$

and then for bx

$$bx = \frac{1}{2a} [-b^2 + b\sqrt{b^2 - 4ac}]$$

multiply by 2 on top and bottom on the right-hand side:

$$bx = \frac{1}{4a} [-2b^2 + 2b\sqrt{b^2 - 4ac}]$$

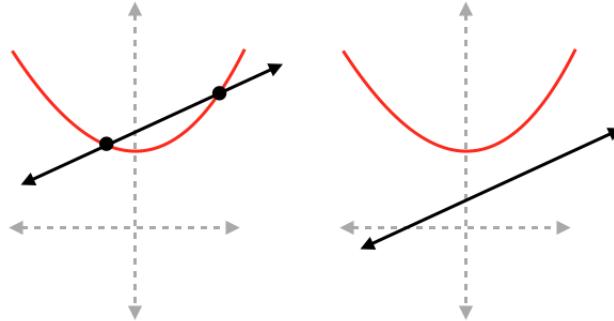
All the terms of bx cancel terms in ax^2 . What's left is $-4ac$ so then

$$\frac{1}{4a} (-4ac) = -c$$

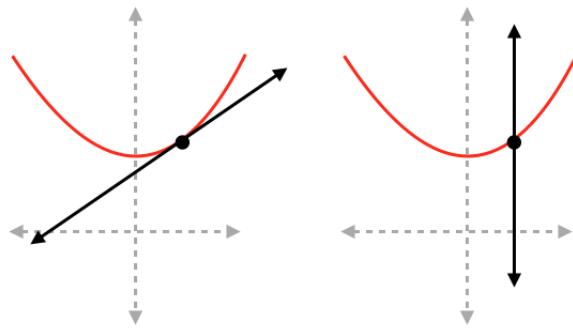
If we had started with the negative root, the square roots in both ax^2 and bx would each change sign, but they would still be opposite signs and hence, cancel.

tangent lines

Consider a parabola and a line on the same graph. There are four possibilities for the intersection of the line and the parabola. First and second: two points (left panel), and no points (right panel).

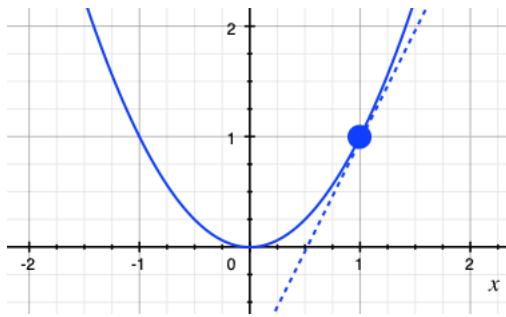


The other two possibilities both have a single point of intersection. The tangent line at a point (left panel), and a vertical line (right panel).



We are particularly interested to know the equations of tangent lines to the parabola, which includes their slopes. Consider the simplest example: $y = x^2$.

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.

The equation of the tangent line is

$$y - y' = m(x - x')$$

Plugging in for $(x', y') = (1, 1)$:

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

$$y = 2x - 1$$

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 "touches" the curve at a single point.

Substitute for y from the equation for the curve:

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

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

Use the quadratic formula to set up an expression for x :

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

There is a single solution when the part under the square root (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.

$y = 2x + y_0$ 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 chord or secant of the parabola, passing through the curve at two points.

The general solution is as follows:

$$y = ax^2$$

$$y = mx + y_0$$

The point(s) of intersection are given by (x, y) satisfying both equations:

$$ax^2 = y = mx + y_0$$

Then

$$ax^2 - mx - y_0 = 0$$

From the quadratic equation

$$x = \frac{m \pm \sqrt{m^2 + 4ay_0}}{2a}$$

For the case of the tangent line, there is a single solution, which happens when the discriminant is zero and then

$$\begin{aligned} x &= \frac{m}{2a} \\ m &= 2ax \end{aligned}$$

As we've been saying. The slope of the tangent to the parabola at a point (x, ax^2) is equal to $2ax$.

We can find the equation of the line by finding y_0 , the value of y when $x = 0$.

$$\begin{aligned} m &= 2ax = \frac{y - y_0}{x - 0} \\ 2ax^2 &= ax^2 - y_0 \\ y_0 &= -ax^2 \end{aligned}$$

Note: do not make the mistake of writing the equation of the line now as

$$y = mx + y_0 = 2ax \cdot x - ax^2$$

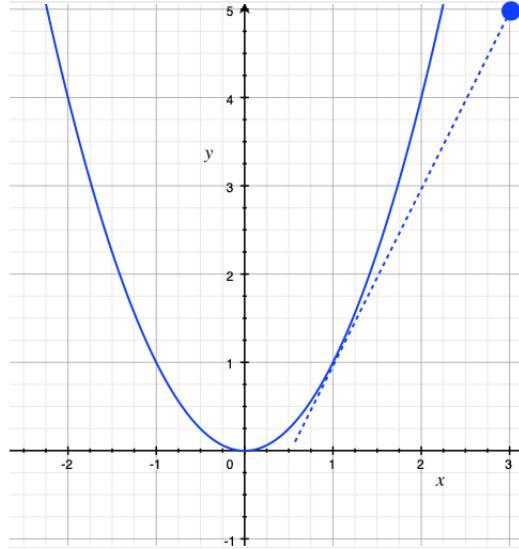
This is wrong because m and y_0 were determined for a particular point on the parabola, but in the equation of the line $y = mx + y_0$, *that* x is any x . Going back to the prime notation we should write:

$$y = mx + y_0 = 2ax' \cdot x - ax'^2$$

where x is a variable and x' is a constant.

tangent passing through a point

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 lines passing through this point with different slopes m are given by:

$$(y - y') = m(x - x')$$

Here, let (x', y') be $(3, 5)$:

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

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. These are points x which are both on the line, and on the curve.

The value of the slope m giving a single solution (zero discriminant) is:

$$(-m)^2 - 4(3m - 5) = 0$$

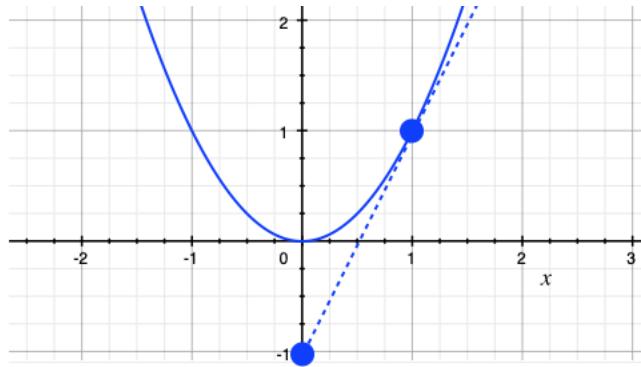
$$m^2 - 12m + 20 = 0$$

Another quadratic. We can factor it by eye:

$$(m - 2)(m - 10) = 0$$

$$m = 2, \quad m = 10$$

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



The original point $(3, 5)$ isn't shown.

As mentioned, there is always another solution which we haven't found explicitly and isn't shown on the graph either. Any vertical line (with infinite slope) passes through only a single point on the parabola.

We ignore this complication.

general solution

We are given the parabola $y = ax^2$, as well as a general point (x', y') , which we suppose is not on the parabola. Any line through the point (x', y') has the equation:

$$y - y' = m(x - x')$$

for all points (x, y) on the line.

If one of those points (x, y) is on the line and *also* on the parabola, it must satisfy $y = ax^2$ as well, so:

$$\begin{aligned} ax^2 - y' &= m(x - x') \\ ax^2 - mx + (mx' - y') &= 0 \end{aligned}$$

The must be only one value of x which works for a particular m , there can be *no other solutions*. (In general, for a point (x', y') *underneath* the parabola there will be two tangent lines).

Write the quadratic equation to solve for x :

$$x = \frac{m \pm \sqrt{m^2 - 4a(mx' - y')}}{2a}$$

There is a single solution when the discriminant is zero. Then

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

$$m = 2ax$$

At the point $(x, y) = (1, 1)$, the slope is equal to 2, if $y = x^2$.

Go back to the equation for the line

$$y - y' = m(x - x')$$

and plug in this slope

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

Then for a given (x', y') we can find the two points (x, y) as

$$ax^2 - y' = 2ax^2 - 2axx'$$

$$ax^2 - 2ax'x + y' = 0$$

$$x = \frac{2ax' \pm \sqrt{(2ax')^2 - 4ay'}}{2a}$$

$$= x' \pm \sqrt{x'^2 - y'/a}$$

There is one solution when the discriminant is zero:

$$y' = ax'^2$$

That's when the given point (x', y') actually was on the curve.

And there are no solutions (in the real numbers) when the discriminant is negative:

$$x'^2 - y'/a < 0$$

$$ax'^2 < y'$$

That is, when the point (x', y') lies in the "cup" of the parabola and $y' > ax'^2$. Then no tangent line runs through (x', y') .

Kline 4-23

The slope of the parabola has some simple interesting properties.

For example, pick two points (x, y) and (x', y') on such that the line joining them goes through the focus.

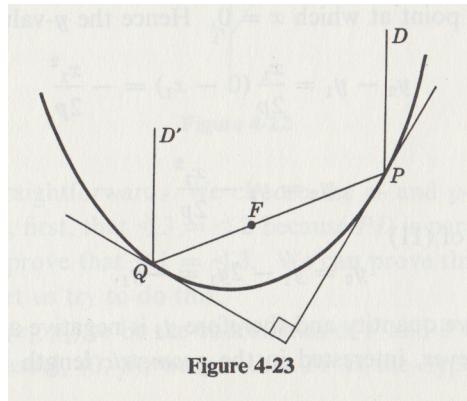


Figure 4-23

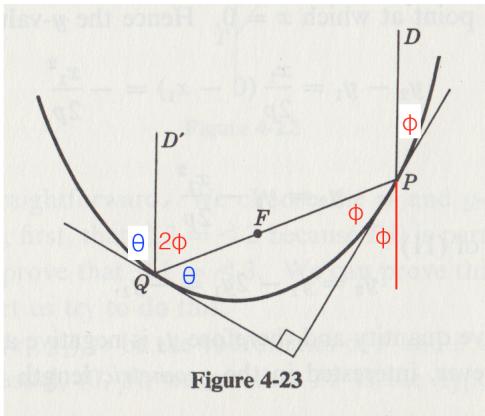
The tangents through these two points form a right angle, as shown in the figure.

Proof.

Recall that when the vertical DP is extended downward, the angle that it makes with the tangent, ϕ , is equal to the angle DP itself makes with the tangent.

Furthermore, FP makes the same angle with the tangent, by the "headlight property". The angle of incidence is equal to the angle of reflection.

The situation is this:



Since DP is parallel to $D'P$, by alternate interior angles the angle $D'QF$ has measure 2ϕ . Using the headlight property again we have that $\theta = \theta$.

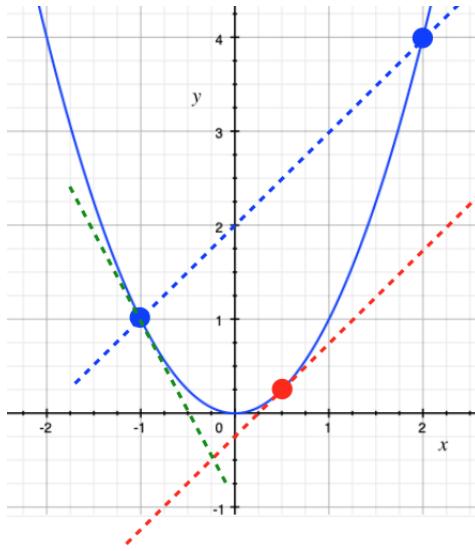
Since $2\phi + 2\theta$ is 180 degrees, θ and ϕ are complementary, so the angle where the tangents meet is a right angle.

□

further comment

Next, 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

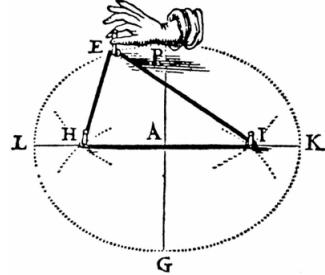
$$\begin{aligned}
 &2a \cdot \frac{1}{2}(x' + x) \\
 &= a(x' + x)
 \end{aligned}$$

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' .

Chapter 6

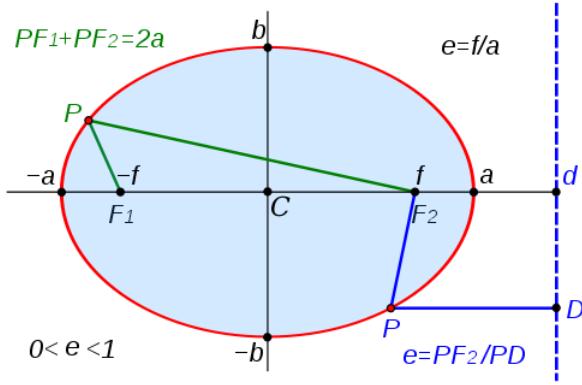
Ellipse

construction



Learning how to draw an ellipse using two pins and a circular piece of string holding a pencil is an early adventure in mathematics. The ellipse is the set of all points whose combined distance to the two pins (foci) is the same.

The drawing is reproduced from a 17th century book in Acheson (see the References).



The pin positions with respect to the origin or center are called the foci, lying at the points shown in the second figure as $(\pm f, 0)$.

We will use the notation c : the focus in the first quadrant is at the point $(c, 0)$.

The lengths of the axes (called semi-major and semi-minor) are usually labeled a and b .

Consider the situation when the pencil is at the point $P = (0, a)$. The distance to the left focus is $c + a$, so the length L of the string is twice that

$$L = 2(c + a)$$

The combined distance from each point on the ellipse to the two foci is the length of the string minus the distance between the two foci

$$L - 2c = 2(c + a) - 2c = 2a$$

standard equation

We learn in algebra that the equation for an ellipse is

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

We will derive this equation below.

The relation of a and c to b can be seen from the point $Q = (0, b)$ (see previous figure) where the combined distance to the two foci is just

$$QF_1 + QF_2$$

From what we said above the distance is $2a$, but Pythagoras also gives us

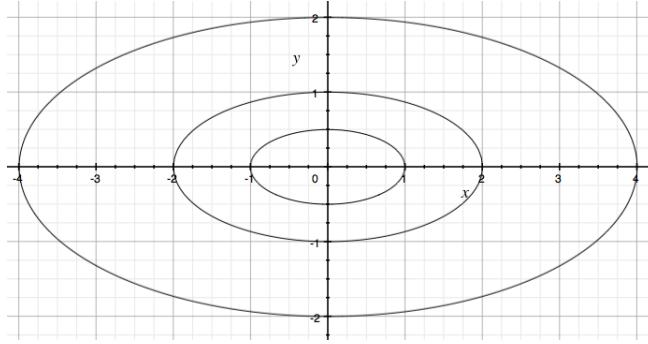
$$QF_1 + QF_2 = 2a = 2\sqrt{b^2 + c^2}$$

so

$$\begin{aligned} b^2 + c^2 &= a^2 \\ c^2 &= a^2 - b^2 \end{aligned}$$

Given a and b one can then find c easily.

Here are three ellipses drawn with the same center.



The difference is an adjustment in the value on the right-hand side of the equation

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

where $r = \{1/2, 1, 2\}$. This is equivalent to scaling both a and b by the same factor of r

$$\frac{x^2}{(ra)^2} + \frac{y^2}{(rb)^2} = 1 = \left(\frac{x/a}{r}\right)^2 + \left(\frac{y/b}{r}\right)^2$$

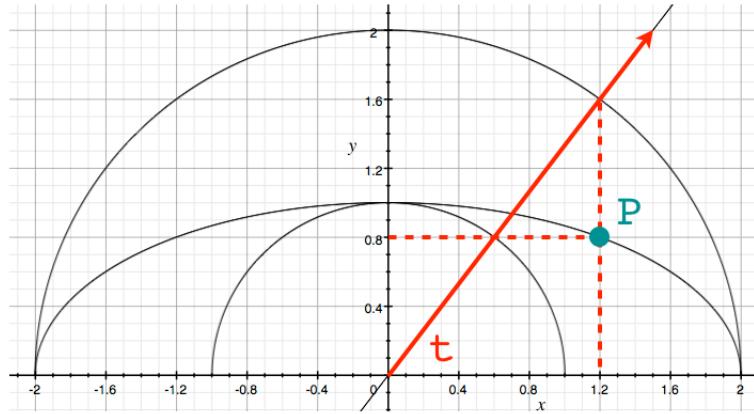
When $r = 2$ we need to make the string a bit less than twice as long, because the length c is also involved:

$$\frac{L_2}{L_1} = \frac{ra + c}{a + c}$$

parametrization

An alternative view is the one below, which shows (black curves) the upper half of two circles of radius $r = 1$ and $r = 2$ and an ellipse whose equation is

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



Here $a = 2$ and $b = 1$.

The standard parametrization of the ellipse is

$$x = a \cos t$$

$$y = b \sin t$$

which I had trouble visualizing, until I drew the picture. The thing is that the parameter t is *not* the angle that a ray to P makes with the x -axis, as it is for the circle. Instead, to find the x value of P corresponding to t , we extend the ray with angle t to the larger circle, with radius a , where we read off the x -value as

$$x = a \cos t$$

We go back to find the intersection of the same ray with the small circle to get

$$y = b \sin t$$

The algebraic way to do this is to show that the parametrization is equivalent to the original formulation

$$x^2 = a^2 \cos^2 t$$

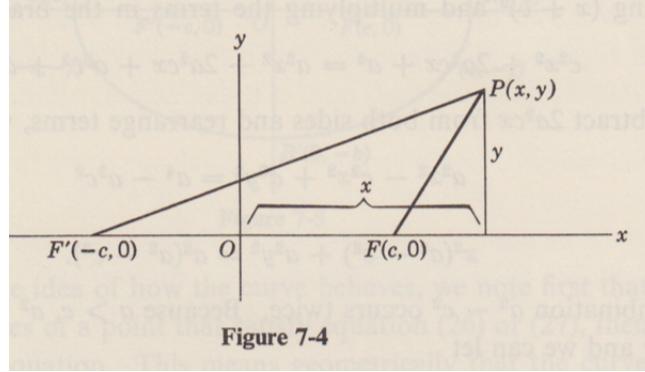
$$y^2 = b^2 \sin^2 t$$

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

as expected.

Derivation of the equation of the ellipse

Although it is a bit tedious, it's a reasonable exercise to derive the equation of the ellipse from the geometric constraint. Recall that a is the length of the semi-major axis and c the distance to each of the foci from the origin.



For any point x, y on the ellipse, the distance to the focus in the first quadrant is

$$\sqrt{(x - c)^2 + y^2}$$

and combined distances to both foci are equal to $2a$ so

$$2a = \sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2}$$

Now we just do some algebra. Pick one square root and rearrange

$$\sqrt{(x - c)^2 + y^2} = 2a - \sqrt{(x + c)^2 + y^2}$$

Square both sides

$$(x - c)^2 + y^2 = 4a^2 - 4a\sqrt{(x + c)^2 + y^2} + (x + c)^2 + y^2$$

Cancel y^2

$$(x - c)^2 = 4a^2 - 4a\sqrt{(x + c)^2 + y^2} + (x + c)^2$$

But

$$(x + c)^2 - (x - c)^2 = 4xc$$

so

$$\begin{aligned} 0 &= 4a^2 - 4a\sqrt{(x + c)^2 + y^2} + 4xc \\ a^2 + xc &= a\sqrt{(x + c)^2 + y^2} \end{aligned}$$

Square again

$$\begin{aligned} a^4 + 2a^2xc + x^2c^2 &= a^2(x^2 + 2xc + c^2 + y^2) \\ a^4 + 2a^2xc + x^2c^2 &= a^2x^2 + 2a^2xc + a^2c^2 + a^2y^2 \end{aligned}$$

Cancel $2a^2xc$

$$a^4 + x^2c^2 = a^2x^2 + a^2c^2 + a^2y^2$$

Gather terms

$$\begin{aligned} a^4 - a^2c^2 &= a^2x^2 - x^2c^2 + a^2y^2 \\ a^2(a^2 - c^2) &= x^2(a^2 - c^2) + a^2y^2 \end{aligned}$$

Recall that $b^2 = a^2 - c^2$

$$b^2a^2 = b^2x^2 + a^2y^2$$

Divide by a^2b^2

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

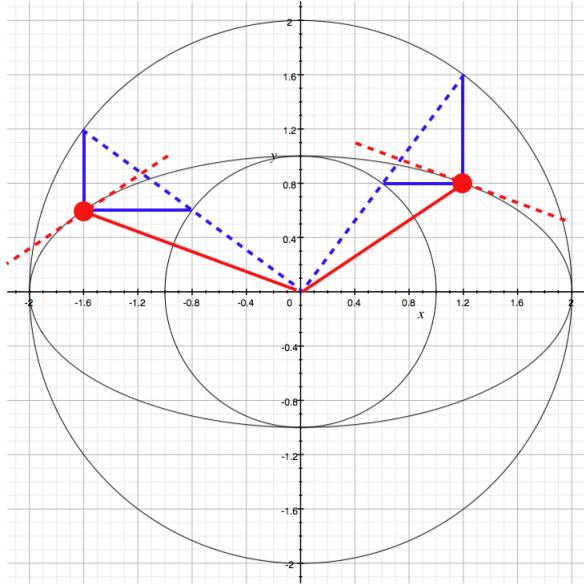
rotation

Let's return to the diagram of the ellipse with two bounding circles of radius a and radius b . There is a new diagram below. Consider the coordinates of the point $P = (x, y)$ (the red dot in the first quadrant) as functions of the angle t . As we said, t is *not* the angle of a ray from the origin to P .

Draw a ray (blue dotted line) from the origin makes an angle t with the x -axis. As before, extend the ray to the outer circle. The radius is a , the angle is t , and

$$a \cos t = x$$

This is the parametrization of the ellipse introduced previously.



The ray drawn with angle t has the same x -intercept with the outer circle as our point P on the ellipse. Similarly, the intercept of the ray with the inner circle has the same y -value as the point P on the ellipse.

We estimate the point $P = (1.2, 0.8) = (6/5, 4/5)$. Using our algebraic equation:

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

Recall that $a = 2$ and $b = 1$ so

$$x^2 + 4y^2 = 4$$

Plugging in for x^2 and y^2 we get

$$\frac{36}{25} + 4 \left(\frac{16}{25} \right) = \frac{100}{25} = 4$$

as expected. Reading off the intercepts for the ray with angle t (dotted blue line) with the outer circle, we have the point $(1.2, 1.6)$ at a distance 2 from the origin. Thus,

$$\frac{1.2}{2} = 0.6 = \cos t$$

$$t \approx 0.927 \text{ rad} \approx 53^\circ$$

Looking again at the figure, we want to consider what happens for the angle $u = t + \pi/2$. This is the dotted blue ray in the second quadrant.

We might calculate the values of sine and cosine for u , but notice that if we view u as a vector, its *dot product* with t must be equal to zero. The coordinates of the intercept of the rotated vector with the outer circle are $(-1.6, 1.2)$, so the cosine of the angle u is

$$\begin{aligned}\cos u &= -0.8 \\ u \approx 2.498 &= t + \frac{\pi}{2} \text{ rad} \approx 143^\circ\end{aligned}$$

We confirm that

$$2.498 - 0.927 = 1.57 = \frac{\pi}{2}$$

The coordinates of the point on the ellipse are $(-1.6, 0.6)$, which we check against the formula

$$\begin{aligned}x^2 + 4y^2 &= 4 \\ (-1.6)^2 + 4(0.6)^2 &= 2.56 + 4(0.36) = 4\end{aligned}$$

(no clean fractions this for this one).

tangent

Finally, and this is really the crucial result:

the vector to the point, call it Q , on the ellipse (red dot in the second quadrant) is the *tangent to the ellipse* for the point P in the first quadrant.

How did this happen? Recall what we did. We had

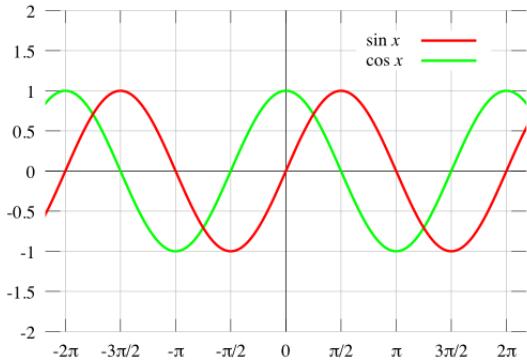
$$x = a \cos t$$

$$y = b \sin t$$

The rotated point $Q = (x', y')$ is

$$x' = a \cos(t + \frac{\pi}{2})$$

$$y' = b \sin(t + \frac{\pi}{2})$$



Sine is like cosine, but shifted to the right by $\pi/2$

$$\cos \theta = \sin(\theta + \frac{\pi}{2})$$

$$\sin \theta = -\cos(\theta + \frac{\pi}{2})$$

So

$$x' = a \cos(t + \frac{\pi}{2}) = -a \sin t$$

$$y' = b \sin(t + \frac{\pi}{2}) = b \cos t$$

Let's look at the position vector, which can be written $\mathbf{r}(t)$, since it's a function of the angle t or the time, but we will just use \mathbf{r} . It has components x and y .

$$\mathbf{r} = \langle x, y \rangle = \langle a \cos t, b \sin t \rangle$$

Now, the tangent to the ellipse is precisely the direction in which a particle at (x, y) is currently moving on the ellipse. The tangent vector points in the same direction as the velocity vector, but \mathbf{v} is just the time-derivative of the position vector.

$$\begin{aligned} \mathbf{v} &= \frac{d\mathbf{r}}{dt} \\ &= \left\langle \frac{dx}{dt}, \frac{dy}{dt} \right\rangle \\ &= \langle -a \sin t, b \cos t \rangle \\ &= \langle x', y' \rangle \end{aligned}$$

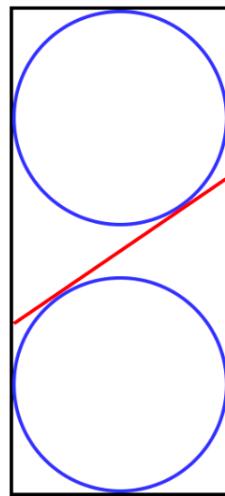
These two methods — using the time-derivative as the tangent, and rotation of t by $\pi/2$ — generate the same vector. And that's the point. :)

Starbird

Here is a neat approach to the ellipse that I saw in one of Michael Starbird's lectures.

Imagine a glass cylinder, shown here in cross-section and colored black. The cylinder has been sliced through at an angle by a plane, and we suppose a flat piece of glass in the shape of an ellipse is glued between the two halves.

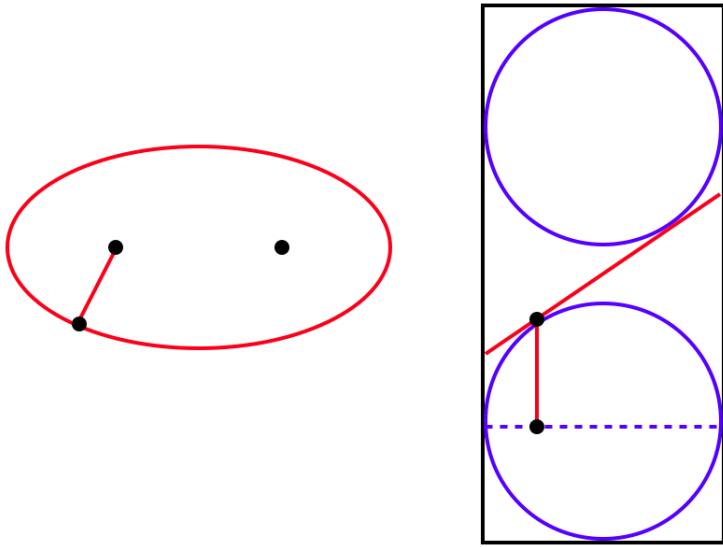
The elongated region in red (formed at the plane of the cut) is the ellipse, and the cylinder is oriented so that at each horizontal position going across the page, the two points on the ellipse are at the same vertical position. We see the plane of the cut edge-on.



Two spheres that fit snugly inside the cylinder lie above and below the ellipse, just touching it. The planar surface of the ellipse is tangent to the spheres, touching each one at a single point.

We claim that the points where the spheres touch the ellipse are the foci of the ellipse.

By the nature of the construction, the two spheres just fit inside the cylinder. That means the intersection where the spheres touch the cylinder is a circle, the lower one is shown with a dotted blue line in the next figure.

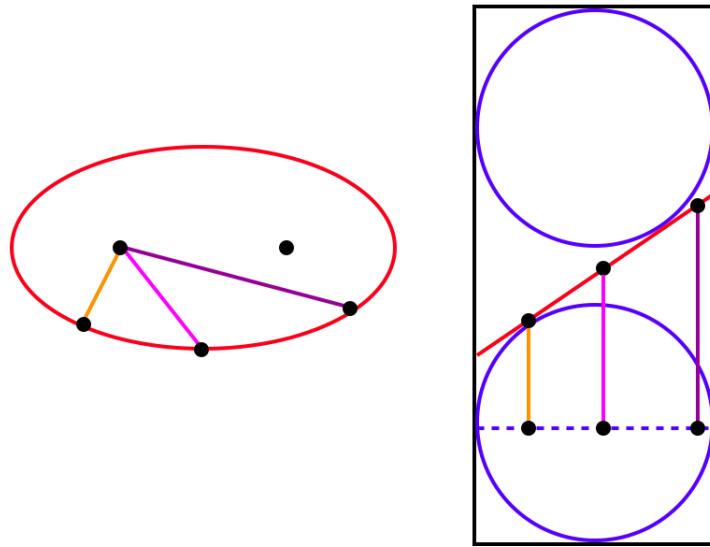


Now consider any point on the ellipse. On the left, we see one point on the ellipse together with two interior points we claim are foci, with a line drawn from our point to one of the foci.

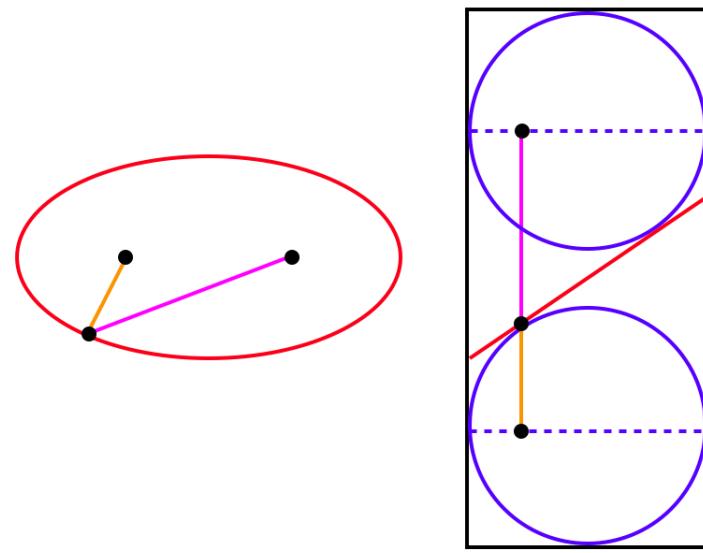
We said that this point is the point where the ellipse touches the lower sphere. We conclude that the line we've drawn from the edge of the ellipse to the focus is a tangent to the sphere.

A second tangent of interest is the perpendicular dropped vertically down the surface of the cylinder, shown in the right panel. Since they are both tangents, this line is the same length as line to the focus.

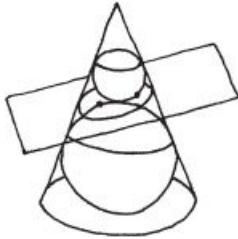
But the construction, and this equality, holds for any point on the ellipse, as shown in the next figure.



Finally, this is true for both spheres (below). The sum of the perpendicular tangents for any point is a constant.



Thus, the points where the spheres touch the ellipse are its foci, because the sum of the distances to any point on the ellipse, which is equal to the sum of the vertical tangents, is a constant.



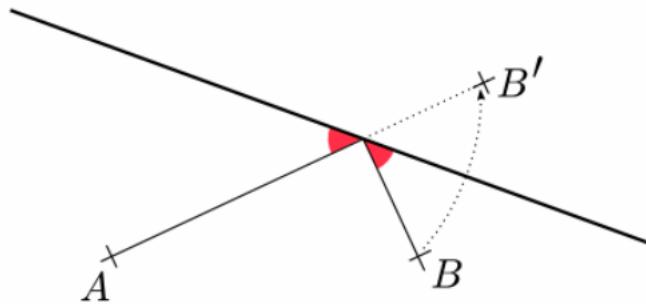
According to Lockhart, the same argument can be used to prove that the cross sections of a cone are ellipses (which seems strange at first since we've been demonstrating that the cross-sections of cylinders are also ellipses).

reflected rays

In any ellipse, the segments from the foci to any point on the ellipse make equal angles with the tangent. This means that light rays emitted from one focus and striking anywhere on the ellipse will pass through the other focus upon reflection. It is the principle behind "whispering galleries."

Here is a simple geometric proof.

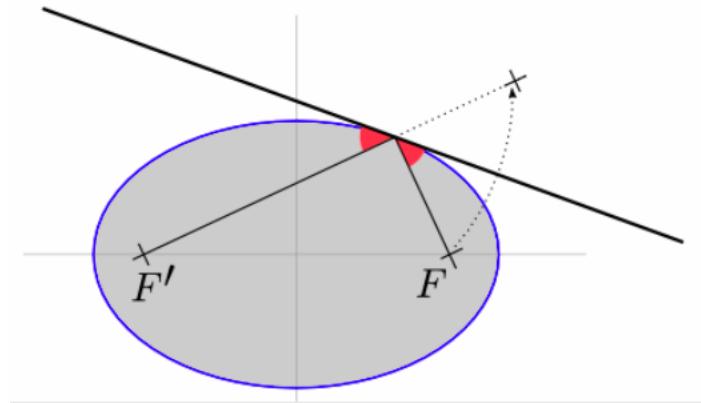
We consider the problem of the "shortest path."



The problem is to go from A to the line and then back to B by the shortest path. The clever solution is to place B' on the other side of the line at the same distance away. By definition (see Euclid) the shortest path A to B' is a straight line.

We can use vertical angles (or supplementary angles twice) and then similar triangles to prove that the two angles colored red are equal.

Now consider an enhanced diagram of the same situation:



We draw the tangent to the ellipse. By definition, the tangent has only a single point on the curve. This point lies at a distance $2a$ from the combined foci. All other points on the line are farther away from the two foci than the point of intersection. (You would have to make the string bigger to draw the ellipse that goes through any of those points).

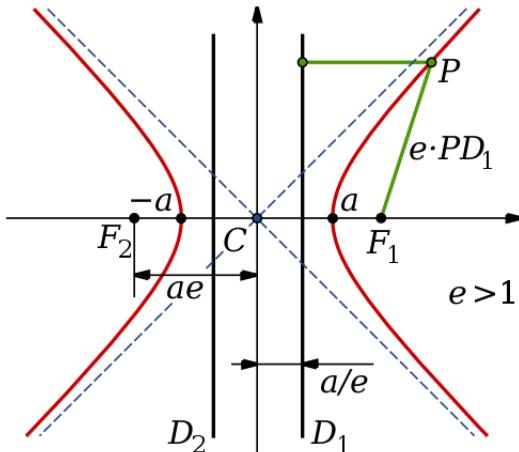
Therefore, the path shown is the shortest path from F' to the tangent and then to F . But we know that for the shortest path the angles colored red are equal.

<http://math.stackexchange.com/questions/1063977/how-to-geometrically-prove-the-focal-property-of-ellipse>

Chapter 7

Hyperbola

Here is a hyperbola as shown in the wikipedia article on the subject.



Hyperbolas of this type (that open "east-west") have equations of the form

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

Rearranging

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

so the minimum value of x occurs when $y = 0$ and $x = a$.

The *conjugate* hyperbola of this one is

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

or equivalently

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

opens "north-south."

And, although I will wait to deal with this complication, we have to mention another very common hyperbola

$$xy = c$$

where it must be true that $x \neq 0$ and $y \neq 0$.

Another feature of hyperbolas is the asymptote, the straight line which is approached when $x, y >> a, b$. In the case of the first example

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

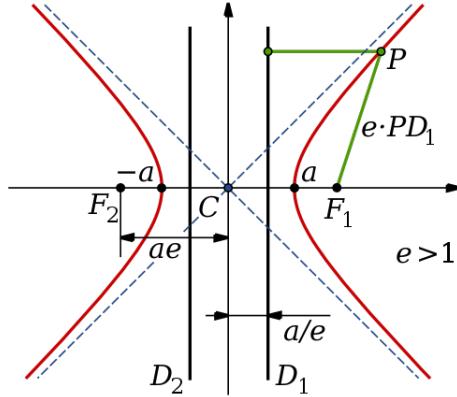
$$y^2 = \frac{b^2}{a^2}x^2 - \frac{1}{a^2}$$

but for large x and y this approaches

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

$$y = \pm \frac{b}{a}x$$

As the diagram suggests:



The following diagram gives geometric meaning to the b coefficient which really derives from the slope of the asymptotic line. We go vertically up from $x = a$ to the asymptote and then go left to the y -axis, that intercept is b .

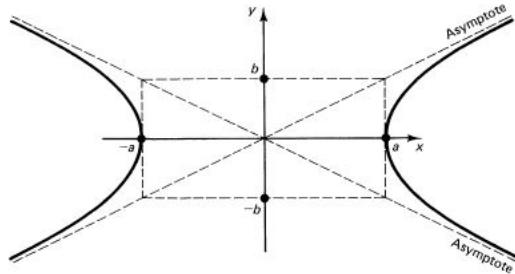
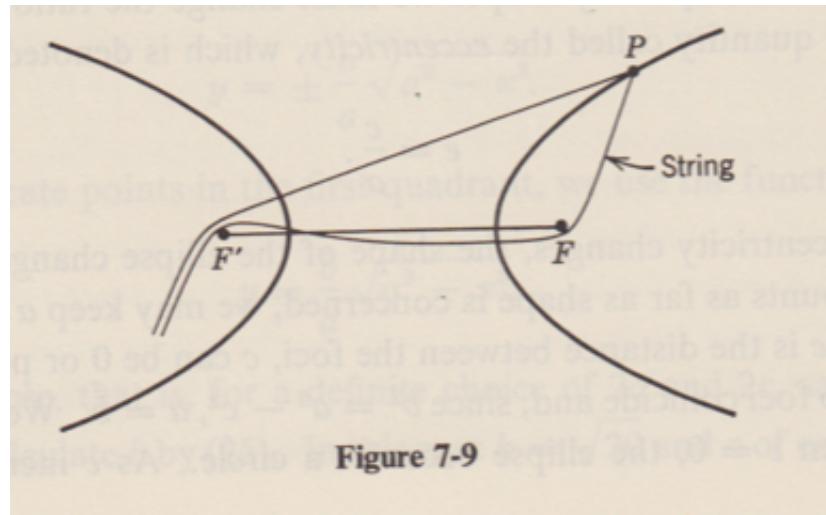


Figure 6.6-1 Hyperbola

geometry

Kline gives the following string and pencil construction for the hyperbola.



Pick two foci F and F' and loop a long piece of string around them, holding it tight. Then place the pencil at some point P on a line between the two foci, at a fixed position in the upper loop.

Now let the string slowly slip up past F' in both directions, increasing the length of PF and PF' by the same amount for each small slip. What this amounts to is that

the difference $PF - PF'$ is constant.

If we place the origin halfway between F and F' then

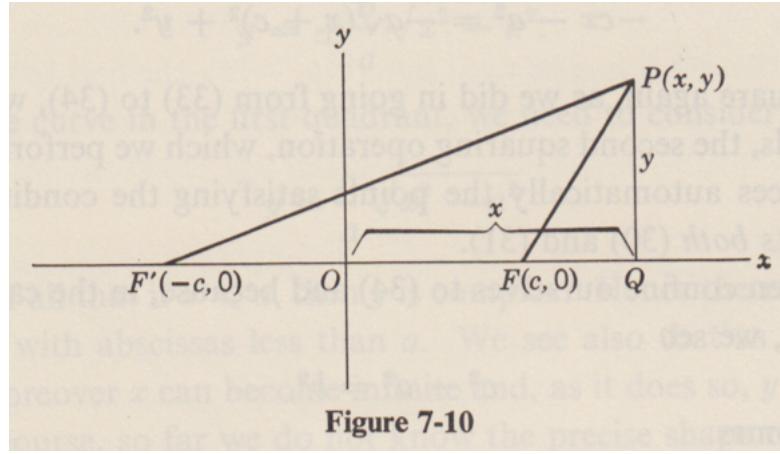


Figure 7-10

$$PF = \sqrt{(x - c)^2 + y^2}$$

$$PF' = \sqrt{(x + c)^2 + y^2}$$

and the difference $PF' - PF$ is

$$\sqrt{(x + c)^2 + y^2} - \sqrt{(x - c)^2 + y^2}$$

and if the constant distance

$$PF' - PF = 2a$$

then

$$\sqrt{(x + c)^2 + y^2} - \sqrt{(x - c)^2 + y^2} = 2a$$

Now we repeat the approach we took for the ellipse:

$$\sqrt{(x + c)^2 + y^2} = 2a + \sqrt{(x - c)^2 + y^2}$$

Square

$$(x + c)^2 + y^2 = 4a^2 + 4a\sqrt{(x + c)^2 + y^2} + (x - c)^2 + y^2$$

Cancel y^2

$$(x + c)^2 = 4a^2 + 4a\sqrt{(x + c)^2 + y^2} + (x - c)^2$$

Since

$$(x + c)^2 - (x - c)^2 = 4cx$$

we have

$$\begin{aligned}4cx &= 4a^2 + 4a\sqrt{(x+c)^2 + y^2} \\cx - a^2 &= a\sqrt{(x+c)^2 + y^2} \\c^2x^2 - 2ca^2x + a^4 &= a^2(x+c)^2 + a^2y^2 \\c^2x^2 - 2ca^2x + a^4 &= a^2x^2 + 2a^2cx + a^2c^2 + a^2y^2 \\(c^2 - a^2)x^2 - a^2y^2 &= (c^2 - a^2)a^2\end{aligned}$$

Define b^2 slightly differently here

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

so

$$\begin{aligned}b^2x^2 - a^2y^2 &= b^2a^2 \\\frac{x^2}{a^2} - \frac{y^2}{b^2} &= 1\end{aligned}$$

which looks familiar.

Part III

Trigonometry

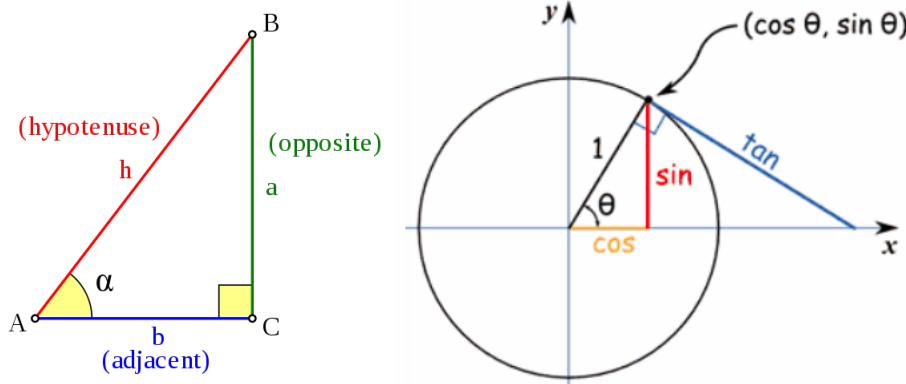
Chapter 8

Six functions

basic definitions

The most elementary trigonometric functions are the sine and cosine. These are defined in geometry as ratios of the lengths of the sides of a right triangle.

Looking at the left panel, we say that the sine of the angle α is the ratio *opposite-over-hypotenuse*, while the cosine of α is the ratio *adjacent-over-hypotenuse*. Tangent is the ratio *opposite-over-adjacent*. The names are abbreviated to three letters in formulas.



Using the notation for the sides from the figure:

$$\sin \alpha = \frac{a}{h}, \quad \cos \alpha = \frac{b}{h}, \quad \tan \alpha = \frac{a}{b} = \frac{\sin \alpha}{\cos \alpha}$$

The "unit circle" is a circle of radius 1 with its center positioned at the origin of coordinates, the place where the x and y axes cross. From the right panel of the diagram you can see that any point (x, y) on the unit circle can be described in radial coordinates as

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

In the diagram, all three right triangles are similar because the red line is an altitude of the largest right triangle. Thus, by similar triangles, the blue side has this relationship

$$\frac{\text{blue side}}{1} = \frac{\sin \theta}{\cos \theta}$$

which explains why it is labeled as $\tan(\alpha)$.

If the vertex labeled B is denoted angle β (the complementary angle of α), then the notions of opposite and adjacent switch so that:

$$\sin \alpha = \cos \beta, \quad \cos \alpha = \sin \beta$$

If the circle has radius r then

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

Stewart:

The mathematicians of ancient India built on the Greek work to make major advances in trigonometry. They [used] the sine (sin) and cosine (cos) functions, which we still do today. Sines first appeared in the Surya Siddhanta, a series of Hindu astronomy texts from about the year 400, and were developed by Aryabhata in Aryabhatiya around 500. Similar ideas evolved independently in China.

The other functions are the inverses of sine, cosine and tangent, namely: cosecant, secant and cotangent. The secant (inverse cosine) comes up sometimes, but the other two are not especially important in calculus.

However, there is one context that we will look at, namely, Archimedes determination of the value of π . The crucial step in that approach will turn out to be the calculation of the cotangent of the half-angle $\theta/2$ given the values of cotangent and cosecant for angle θ .

The main relationship or identity is derived from the Pythagorean theorem. We had above that for a unit circle

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

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

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

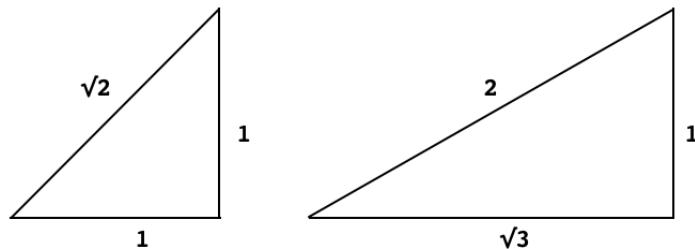
and transformed (dividing by the cosine squared) to

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

particular values

We can easily determine the values for these functions for three special cases.

The first is the angle 45 degrees or $\pi/4$. Draw an isosceles right triangle with sides of length 1 (left panel).



Then the hypotenuse has length $\sqrt{2}$ (from Pythagoras) and the values are

$$\sin \frac{\pi}{4} = \frac{1}{\sqrt{2}} = \cos \frac{\pi}{4}$$

$$\tan \frac{\pi}{4} = 1$$

For the other two, bisect an equilateral triangle and erase one half (right panel). The smaller angle is 30 degrees or $\pi/6$ and its complement is 60 degrees or $\pi/3$.

The values are

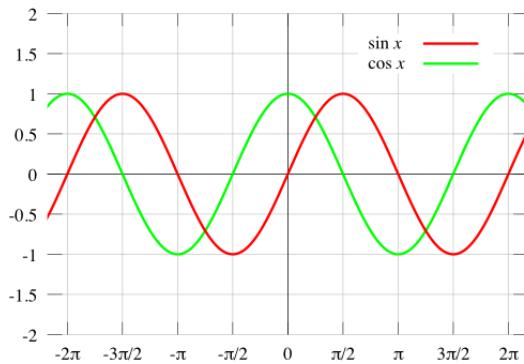
$$\sin \frac{\pi}{6} = \frac{1}{2} = \cos \frac{\pi}{3}, \quad \cos \frac{\pi}{6} = \frac{\sqrt{3}}{2} = \sin \frac{\pi}{3}$$

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

We can verify that

$$\left(\frac{1}{\sqrt{2}}\right)^2 + \left(\frac{1}{\sqrt{2}}\right)^2 = 1, \quad \frac{1}{2}^2 + \left(\frac{\sqrt{3}}{2}\right)^2 = 1$$

graph



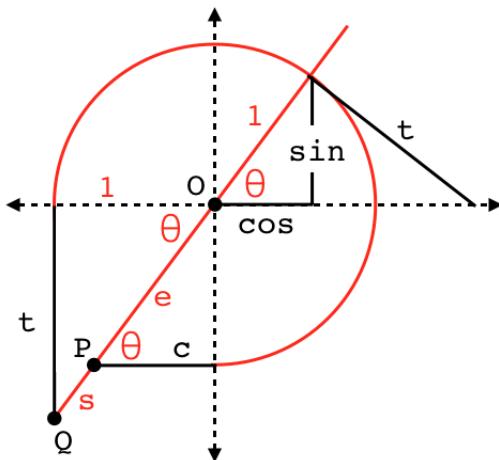
Savov:

The sine function represents a fundamental unit of vibration. The graph of $\sin(x)$ oscillates up and down and crosses the x -axis multiple times. The shape of the graph of $\sin(x)$ corresponds to the shape of a vibrating string.

Imagine a circle placed to the left of a graph. We can think of the sine function as the vertical "shadow" of the y -value of the point (x, y) as it travels around the circle at the same constant speed as the point on the graph "moves" to the right. Similarly, the cosine is the shadow of the x -value (on the x -axis).

visualization of all six functions

Let us work with a unit circle. Draw the radius to form the angle θ and then draw the vertical and horizontal components which we know are $\sin \theta$ and $\cos \theta$.



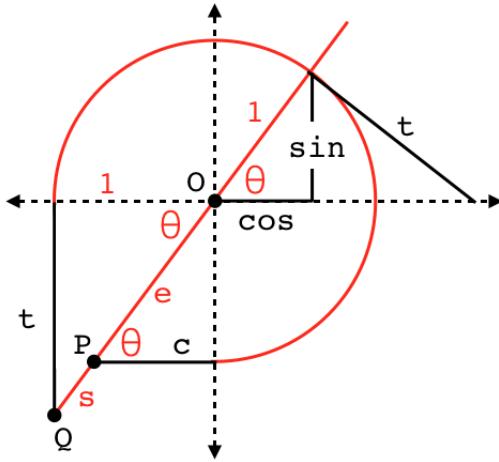
By extending the radius, we can draw several other triangles which are similar to the original one. The angle equalities follow from the vertical angle and alternate interior angle theorems.

In the first quadrant, the similar triangle with opposite side t has a tangent which is $t/1$, but by similar triangles this is just $\tan \theta$. It is not an accident that this is also the tangent to the circle at the point where the radius meets it.

A similar argument justifies the label t on the triangle in the third quadrant.

If we label the origin as O and consider the line segment OQ and call that length s , we have that $1/s = \cos \theta$, so $s = \sec \theta$.

There are two more.



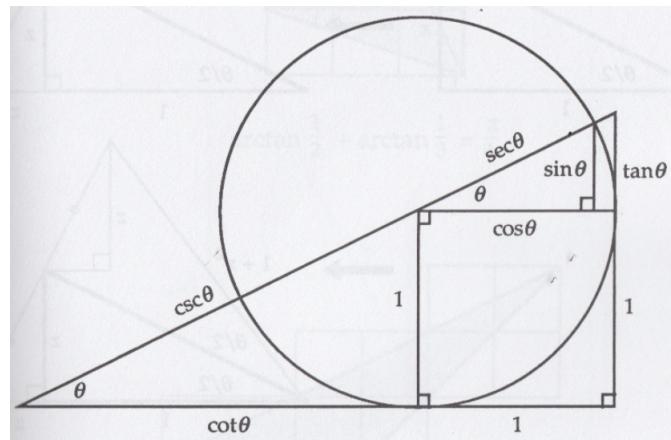
The horizontal length c in the third quarter is drawn horizontally to meet the y -axis at the same point where the circle crosses. c is the adjacent side for angle θ and the ratio with the opposite side which is just the radius of the circle is c , but this is also the inverse of the tangent, the cotangent, so $c = \cot \theta$.

Finally, if e is the length of the line segment OP , then

$$\cos \theta = \frac{c}{e} = \frac{\cot \theta}{e} = \frac{\cos \theta}{\sin \theta} = \frac{1}{e}$$

We have that e is the inverse sine, known as the cosecant, and $e = \csc \theta$.

Here is another version that is similar, from Proofs without words:



This gives various identities simply, like

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

Our source points out that we can also obtain:

$$(1 + \tan \theta)^2 + (1 + \cot \theta)^2 = (\sec \theta + \csc \theta)^2$$

I've certainly never seen that last one, but we can try to prove it algebraically in reverse.

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

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

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

And $1 + \cot^2 \theta = \csc^2 \theta$ so:

$$2 \tan \theta + 2 \cot \theta = 2 \sec \theta \csc \theta$$

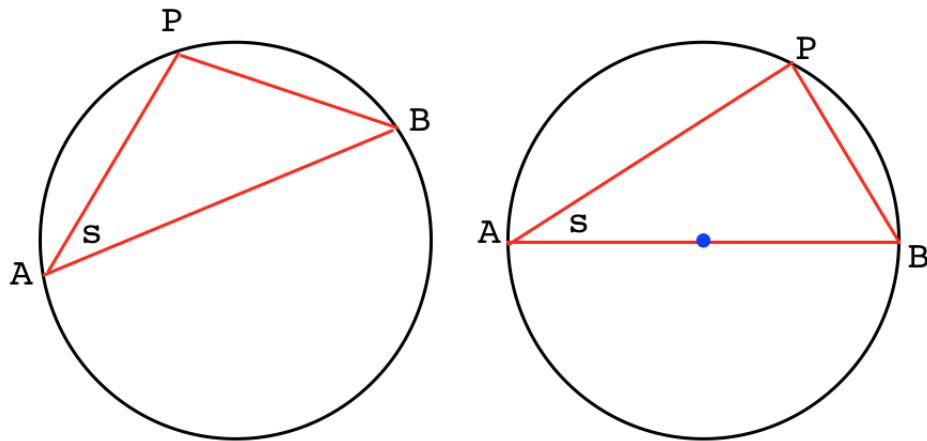
Multiply by $\sin \theta \cos \theta$:

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

which is correct.

chords of a circle

In our work on arcs and chords of a circle, we found that equal angles on the perimeter of the circle subtend equal arcs. So, for example, in this figure the angle s subtends the same arc in both panels, and a chord of the same length, regardless of where it intersects the perimeter.



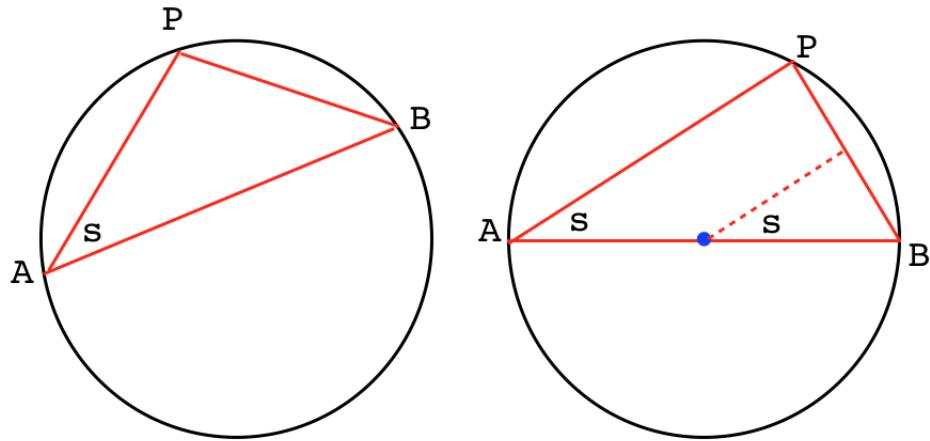
We also showed that, as in the right panel, the angle APB is a right angle, since AB is a diameter of the circle.

Therefore, PB is the opposite side in a right triangle. As a result:

$$\frac{PB}{2r} = \sin s$$

and this is true whether or not one of the sides flanking s is a diameter of the circle.

A second proof of this is the following: erect the perpendicular bisector of PB .



The small triangle containing the dotted line as one edge and the vertex at B is similar to the large triangle ABP , because the angle at the vertex B is shared and the two sides PB and AB are in proportion.

Therefore, the small triangle is a right triangle, the central angle is s , and now one-half of PB divided by r is the sine of s .

$$\sin s = \frac{PB/2}{r} = \frac{PB}{2r}$$

□

Chapter 9

Sum of angles

There are some really important formulas that relate the sine and cosine of individual angles to the sine and cosine of their sum (or difference). Here is one of them. For angles s and t

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

By $\cos s - t$ we mean $\cos(s - t)$, but have left off the parentheses.

There are four formulas, and then some special examples. These are used a lot in calculus, not only for solving problems, but most important, in finding an expression for the derivatives of the sine and cosine functions.

You really must know them. I think it's so important that we will show several ways of finding them.

The proofs are also beautiful, which helps to explain why I've included so many. The easiest way to remember them uses Euler's equation, but since you've probably never seen that, I'll put it off until the end.

I've memorized only the one given above. Say "cos cos" and then recall the difference in sign, minus on the left, plus on the right.

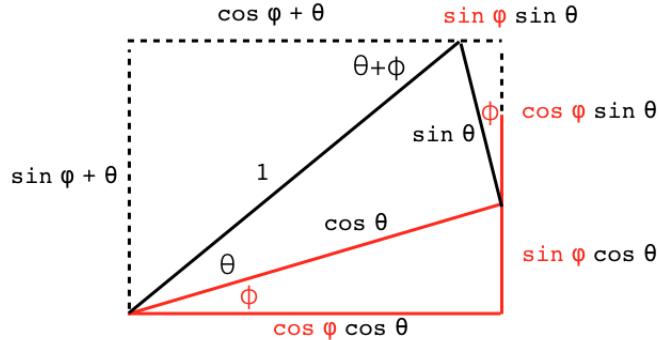
I like this version because it can be checked easily. Just set $s = t$:

$$\cos s - t = \cos s - s = \cos 0 = 1 = \cos^2 s + \sin^2 s$$

which is our favorite trigonometric identity and obviously correct.

similar triangles

Draw two triangles, one on top of the other, with the hypotenuse of the second scaled to be equal to 1. Then draw a rectangle around the whole thing.



For the triangle with angle θ and hypotenuse 1, the labels should be obvious.

Second, for triangles with angle ϕ where the hypotenuse is *not* 1, we have something like $\cos\phi\cos\theta$ on the bottom of the figure, which gives the desired value $\cos\phi$ after dividing by the hypotenuse, $\cos\theta$.

The angle labeled $\theta + \phi$ at the top is known by the alternate interior angles theorem, and the angle ϕ at top right is by complementary and supplementary angles.

Now, just read off the relationships from the sides of the rectangle:

$$\sin\phi + \theta = \sin\phi\cos\theta + \cos\phi\sin\theta$$

$$\cos\phi + \theta = \cos\phi\cos\theta - \sin\phi\sin\theta$$

change signs

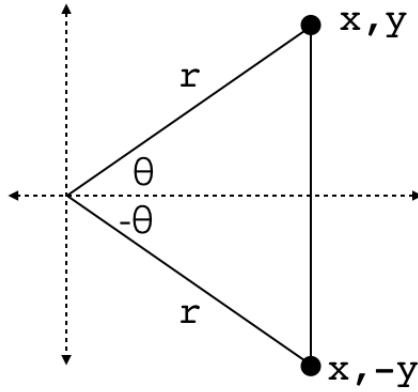
For $\cos s - t$, flip the sign on the second term.

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

That's because

$$\cos -\theta = \cos\theta$$

$$\sin -\theta = -\sin\theta$$



The diagram shows the reason:

$$\cos \theta = x/r = \cos -\theta$$

while

$$\sin \theta = y/r = -(\sin -\theta) = -(-y/r)$$

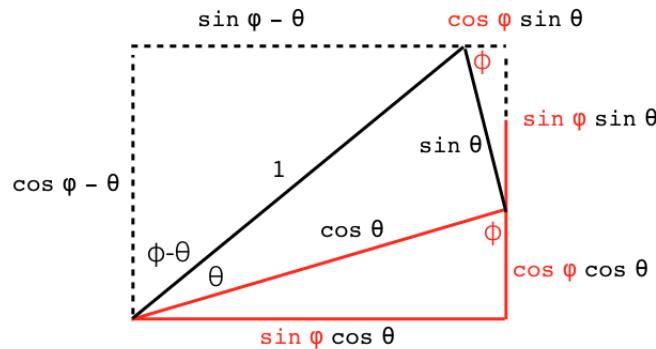
Substitute $-\sin \theta$ for $\sin -\theta$ and $\cos \theta$ for $\cos -\theta$:

$$\begin{aligned} \cos s - t &= \cos s \cos -t + \sin s \sin -t \\ &= \cos s \cos t - \sin s \sin t \end{aligned}$$

and

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

It's kind of overkill, but still worth noting that a simple change to the figure we had above will give the difference formulas:



We've changed the symbol ϕ to refer to the complementary angle from what it was before.

We can justify the label $\phi - \theta$ for the angle at the lower left in various ways, for example, by adding up the three angles at that corner:

$$(\phi - \theta) + \theta + (90 - \phi) = 90$$

Switch the labels appropriately (it's easy since this ϕ is the complement of the old one).

Read the result:

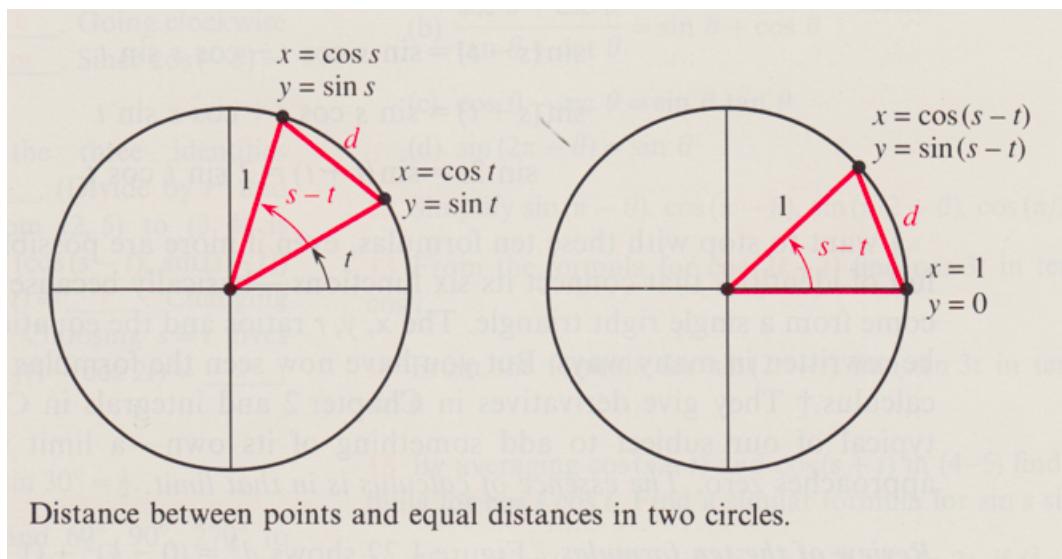
$$\sin \phi - \theta = \sin \phi \cos \theta - \cos \phi \sin \theta$$

$$\cos \phi - \theta = \cos \phi \cos \theta + \sin \phi \sin \theta$$

Here are several other derivations that I've come across over the years. You only need what we've given above, but they are interesting to work through and give practice in dealing with trig functions.

Strang

For a geometric derivation of the sum of angles formula with minimal setup, I really like this figure from Strang



We have the same triangle in the two panels, just rotated clockwise on the right.

The squared distance between two points in the plane is

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 = \Delta x^2 + \Delta y^2$$

This is just the Pythagorean theorem in disguise.

In the left panel, t is the angle between the lower radius and the x -axis, s is the angle between the upper radius and the x -axis, and as labeled, $s - t$ is the angle between the two radii.

The distance d squared for the two points on the circle in the left panel is

$$d^2 = (\cos s - \cos t)^2 + (\sin s - \sin t)^2$$

Multiply out:

$$d^2 = \cos^2 s - 2 \cos s \cos t + \cos^2 t + \sin^2 s - 2 \sin s \sin t + \sin^2 t$$

We have two copies of $\sin^2 + \cos^2$, one for angle s and one for angle t

$$d^2 = 2 - 2 \cos s \cos t - 2 \sin s \sin t$$

In the right panel, the two radii have been rotated, preserving the same angle between them.

$$d^2 = (\cos(s - t) - 1)^2 + \sin(s - t)^2$$

(Don't forget the 1).

$$\begin{aligned} &= \cos^2(s - t) - 2 \cos(s - t) + 1 + \sin^2(s - t) \\ &= 2 - 2 \cos(s - t) \end{aligned}$$

Because the included angle hasn't changed, neither has the distance, so we can equate the two expressions.

$$2 - 2 \cos(s - t) = 2 - 2 \cos s \cos t - 2 \sin s \sin t$$

Subtract 2 from both sides, divide by 2, and change all the signs leaving

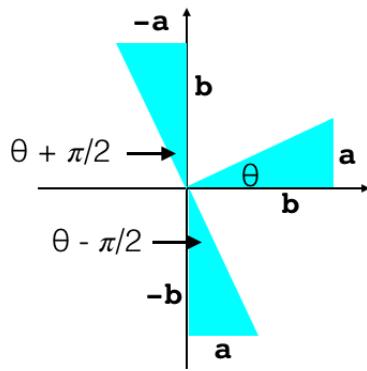
$$\cos(s - t) = \cos s \cos t + \sin s \sin t$$

This is our formula for the cosine of the difference of two angles.

getting to sine

Strang's derivation gives us the formula for sum and difference of cosines. To get the formula for the sine in the same way, we would need to mix sine and cosine when computing a distance in his diagram. I don't know how to do that. So our problem is how to go from the cosine formula to the sine formula.

Let's look at the relationships between sine and cosine for angles that are related by addition or subtraction of $\pi/2$.



In the figure, I have simply rotated the same triangle.

What we see is that

$$\sin \theta + \frac{\pi}{2} = b = \cos \theta$$

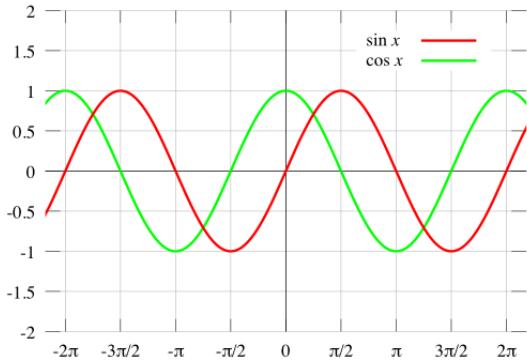
$$\cos \theta + \frac{\pi}{2} = -a = -\sin \theta$$

and

$$\sin \theta - \frac{\pi}{2} = -b = -\cos \theta$$

$$\cos \theta - \frac{\pi}{2} = a = \sin \theta$$

Here is an alternative connection which proceeds from the graph of sine and cosine versus the angle.



so it's easy to see that $\cos t = \sin t + \pi/2$.

So then, consider

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

Suppose we modify the left-hand side to be

$$\cos s + t - \frac{\pi}{2}$$

that gives

$$\cos(s + t - \frac{\pi}{2}) = \cos s \cos(t - \frac{\pi}{2}) - \sin s \sin(t - \frac{\pi}{2})$$

If you go back to our table above and substitute for $\cos(t - \frac{\pi}{2}) = \sin t$

$$\cos(s + t - \frac{\pi}{2}) = \cos s \sin t - \sin s \sin(t - \frac{\pi}{2})$$

then for $\sin(t - \frac{\pi}{2}) = -\cos t$

$$\cos(s + t - \frac{\pi}{2}) = \cos s \sin t + \sin s \cos t$$

and then finally for $\cos(s + t - \frac{\pi}{2}) = \sin s + t$

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

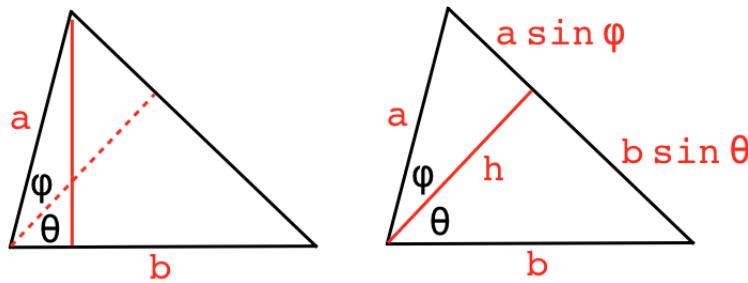
Most people find this difficult to do without making a mistake.

sine of the sum

Here is a very nice geometrical derivation for the sine of the sum. If you like Strang's derivation for the cosine (as I do), this would be a good complement.

We compute the area of the triangle in two different ways. On the left we have that

$$A = \frac{1}{2}ab \sin(\theta + \phi)$$



On the right, the two smaller triangles are right triangles with $h = a \cos \phi = b \cos \theta$.

$$A_{top} = \frac{1}{2}a \sin \phi b \cos \theta$$

$$A_{bottom} = \frac{1}{2}b \sin \theta a \cos \phi$$

These two expressions add to give the total area. We can factor out the common $ab/2$ and write the equality:

$$A = \sin(\theta + \phi) = \sin \phi \cos \theta + \sin \theta \cos \phi$$

<https://www.cut-the-knot.org/triangle/SinCosFormula.shtml>

Euler

Euler's formula is:

$$e^{i\theta} = \cos \theta + i \sin \theta$$

If you've never seen it before, don't worry what it means or where it comes from. Just treat i as a constant with $i^2 = -1$. Multiply as follows:

$$(\cos s + i \sin s)(\cos t + i \sin t)$$

$$\begin{aligned}
&= \cos s \cos t + i^2 \sin s \sin t + i [\sin s \cos t + \cos s \sin t] \\
&= \cos s \cos t - \sin s \sin t + i [\sin s \cos t + \cos s \sin t]
\end{aligned}$$

This is a *complex* number with a real part (the first two terms), plus an imaginary part, the last two terms, with a leading factor of i .

For the same calculation with the exponential

$$\begin{aligned}
e^{is} \cdot e^{it} &= e^{i(s+t)} \\
&= \cos(s+t) + i \sin(s+t)
\end{aligned}$$

By Euler's formula, these two expressions are equal.

The rule for equality of complex numbers is that both the real parts and the imaginary parts must be equal. So we have

$$\begin{aligned}
\cos(s+t) &= \cos s \cos t - \sin s \sin t \\
\sin(s+t) &= \sin s \cos t + \cos s \sin t
\end{aligned}$$

Chapter 10

Double angle

To review very quickly, sine:

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

$$\sin 2t = 2 \sin t \cos t$$

And cosine:

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

$$\cos 2t = \cos^2 t - \sin^2 t$$

$$= 2 \cos^2 t - 1$$

Sometimes it is helpful to have a simpler notation, especially when we want to do algebra. Let's use the symbols S , C , and T for sine, cosine and tangent, respectively.

Mark the values we are deriving (for one-half a double angle) with primes. We have:

$$S = 2S'C'$$

$$C = C'^2 - S'^2 = 2C'^2 - 1$$

The inverse double tangent is then

$$\frac{1}{T} = \frac{2C'^2 - 1}{2S'C'}$$

$$= \frac{1}{T'} - \frac{1}{S}$$

$$\frac{1}{T'} = \frac{1}{T} + \frac{1}{S}$$

This is the formula used to such great effect by Archimedes in obtaining an approximation for π .

$$\cot t = \cot 2t + \csc 2t$$

Another formula that is often included in tables is one for the double tangent. We derive this by going back to the original addition formulas:

$$\begin{aligned}\tan s + t &= \frac{\sin s + t}{\cos s + t} \\ &= \frac{\sin s \cos t + \cos s \sin t}{\cos s \cos t - \sin s \sin t}\end{aligned}$$

Divide through by $\cos s \cos t$:

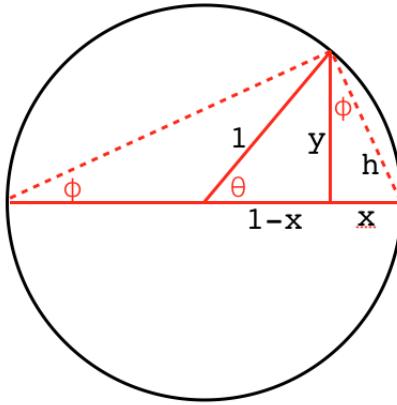
$$= \frac{\tan s + \tan t}{1 - \tan s \tan t}$$

For the double-angle, this becomes:

$$\tan 2t = \frac{2 \tan t}{1 - \tan^2 t}$$

geometric derivation of double angle

Here is a simple geometric derivation of the double angle formula for sine. We start with an inspired diagram.



From our work with arcs, we know that angle ϕ on the left is one-half the central angle θ , and from Thales' theorem, that the angle on the circle at the top-right (formed by the dotted lines) is a right angle.

So the small right triangle with hypotenuse h also has angle ϕ , as labeled, since they have the same complementary angle.

It helps to know where we're going, as well. From above:

$$\sin \theta = 2 \sin \phi \cos \phi$$

Just reading off the small triangle we have that

$$\begin{aligned} \sin \phi \cos \phi &= \frac{x}{h} \cdot \frac{y}{h} \\ &= \frac{xy}{h^2} \end{aligned}$$

and Pythagoras says that $h^2 = x^2 + y^2$ so:

$$= \frac{xy}{x^2 + y^2}$$

We're looking to involve θ . From the right triangle containing that angle:

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

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

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

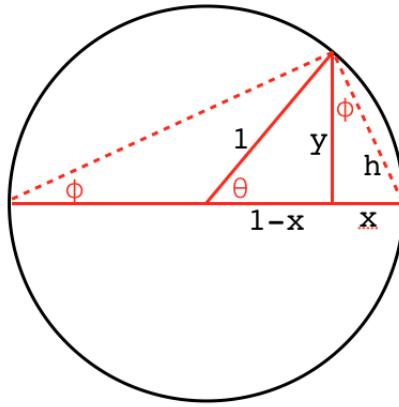
So, substituting into what we had above:

$$\sin \phi \cos \phi = \frac{xy}{2x}$$

$$2 \sin \phi \cos \phi = y = \sin \theta$$

For the other one, again we remind ourselves of the target:

$$\cos \theta = \cos^2 \phi - \sin^2 \phi$$



Substituting, using the diagram:

$$\begin{aligned} \cos \theta &= \frac{y^2}{h^2} - \frac{x^2}{h^2} \\ &= \frac{y^2 - x^2}{h^2} \\ &= \frac{y^2 - x^2}{x^2 + y^2} \end{aligned}$$

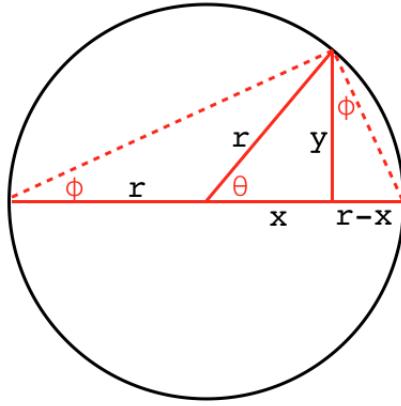
But $2x = x^2 + y^2$:

$$\begin{aligned} &= \frac{2x - x^2 - x^2}{2x} \\ &= 1 - x \end{aligned}$$

$$= \cos \theta$$

□

Similar calculations can be done for a diagram that is slightly relabeled, to provide a simple geometric proof of the Pythagorean theorem.



Thales theorem allows us to deduce that the triangle with two dotted sides is a right triangle, hence the two angles labeled ϕ are equal by complementarity with the same angle. By similar triangles, we have then that

$$\tan \phi = \frac{y}{r+x} = \frac{r-x}{y}$$

$$y^2 = r^2 - x^2$$

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

For the double-angle calculations, there are two triangles with base angle ϕ and the hypotenuse a dotted line. The squared lengths of the two hypotenuses are:

$$\begin{aligned} (r+x)^2 + y^2 &= r^2 + 2rx + x^2 + y^2 \\ &= 2r^2 + 2rx = 2r(r+x) \end{aligned}$$

and

$$\begin{aligned} (r-x)^2 + y^2 &= r^2 - 2rx + x^2 + y^2 \\ &= 2r^2 - 2rx = 2r(r-x) \end{aligned}$$

So then the products of sin and cos work out to be simple ratios, which in each case will have further cancelations.

As one example, for the large triangle:

$$\cos^2 \phi - \sin^2 \phi = \frac{(x+r)^2}{2r(r+x)} - \frac{y^2}{2r(r+x)}$$

Let us just work with the numerator for a minute:

$$\begin{aligned} (x+r)^2 - y^2 &= x^2 + 2xr + r^2 - (r^2 - x^2) \\ 2x^2 + 2xr &= 2x(x+r) \end{aligned}$$

So we see that $2(x+r)$ cancels and the ratio is just

$$\frac{x}{r}$$

which is $\cos \theta$.

another calculation

We found previously that

$$\begin{aligned} \sin \frac{\pi}{4} &= \cos \frac{\pi}{4} = \frac{1}{\sqrt{2}} \\ \sin \frac{\pi}{6} &= \cos \frac{\pi}{3} = \frac{1}{2} \\ \sin \frac{\pi}{3} &= \cos \frac{\pi}{6} = \frac{\sqrt{3}}{2} \end{aligned}$$

These angles correspond to 30, 45 and 60 degrees. It might be nice to have sine and cosine of 15 and 75 degrees as well. That would make even divisions of the first 90 degrees. We can get them as the sum and difference of $\pi/4$ and $\pi/6$.

Let $s = \pi/4$ and $t = \pi/6$. Then

$$\begin{aligned} \sin \frac{\pi}{12} &= \sin s - t = \sin s \cos t - \sin t \cos s \\ &= \frac{1}{\sqrt{2}} \cdot \frac{\sqrt{3}}{2} - \frac{1}{2} \cdot \frac{1}{\sqrt{2}} = \frac{\sqrt{3}-1}{2\sqrt{2}} \end{aligned}$$

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

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

$$\begin{aligned}&\frac{(\sqrt{3}-1)^2 + (\sqrt{3}+1)^2}{(2\sqrt{2})^2} \\ &= \frac{3-2\sqrt{3}+1+3+2\sqrt{3}+1}{8} = 1\end{aligned}$$

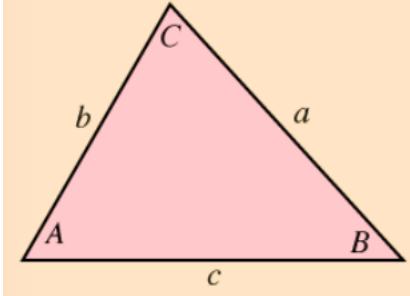
We can calculate similarly for $s+t = 5\pi/12$ or just switch sine and cosine from $\pi/12$.

Chapter 11

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

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


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

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.

If $\angle C$ were a right angle then we would have that

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

but since it's not, there is a correction term

$$c^2 = a^2 + b^2 - \Delta$$

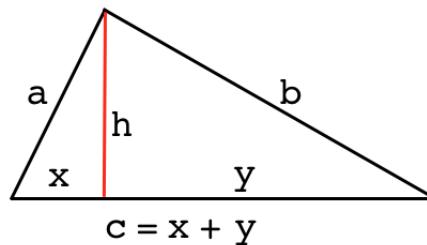
Δ is *subtracted* from the length of c . Δ should be larger, the smaller C gets (as the cosine does), since the opposite side gets squeezed. And it should be larger as the sides a and b are larger, since there is a bigger effect on c in absolute terms.

That's exactly what the law of cosines does! Δ is some factor times a times b times $\cos C$ and the whole term has a minus sign.

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 $c = x + y$ using the vertical altitude from vertex C .



Proof.

$$a^2 - x^2 = h^2$$

$$b^2 - y^2 = h^2$$

So

$$a^2 = x^2 + h^2 = x^2 + b^2 - y^2$$

But

$$y = c - x$$

$$y^2 = c^2 - 2cx + x^2$$

Therefore:

$$\begin{aligned} a^2 &= x^2 + b^2 - (c^2 - 2cx + x^2) \\ a^2 &= b^2 - c^2 + 2cx \end{aligned}$$

Finally, $x = a \cos B$ so

$$\begin{aligned} a^2 &= b^2 - c^2 + 2ac \cos B \\ b^2 &= a^2 + c^2 - 2ac \cos B \end{aligned}$$

This is the law of cosines.

□

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

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab \cos C \\ a^2 &= b^2 + c^2 - 2bc \cos A \end{aligned}$$

Alternate Proof.

Add the first two equations and rearrange:

$$a^2 + b^2 = x^2 + y^2 + 2h^2$$

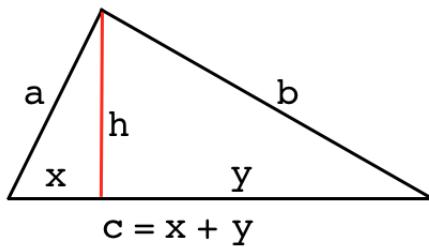
but

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

so

$$\begin{aligned} a^2 + b^2 &= c^2 - 2xy + 2h^2 \\ c^2 &= a^2 + b^2 + 2(xy - h^2) \\ &= a^2 + b^2 - 2(h^2 - xy) \end{aligned}$$

We need to show that $h^2 - xy$ somehow equals $ab \cos C$.



Let the smaller right triangle with hypotenuse a include angle C' , and the one with hypotenuse b have angle C'' , where $C = C' + C''$.

$$\begin{aligned} h &= a \cos C' \\ x &= a \sin C' \\ h &= b \cos C'' \\ y &= b \sin C'' \end{aligned}$$

This reminds us of the sum of cosines:

$$\cos C = \cos C' \cos C'' - \sin C' \sin C''$$

Let's see:

$$\begin{aligned} h^2 &= ab \cos C' \cos C'' \\ xy &= ab \sin C' \sin C'' \end{aligned}$$

So

$$\begin{aligned} h^2 - xy &= ab(\cos C' \cos C'' - \sin C' \sin C'') \\ &= ab \cos C \end{aligned}$$

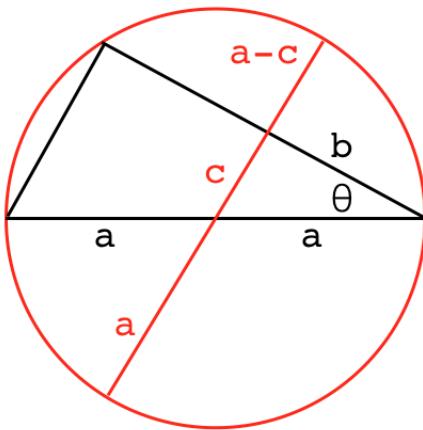
Substituting into the equation we had above:

$$\begin{aligned} c^2 &= a^2 + b^2 - 2(h^2 - xy) \\ &= a^2 + b^2 - 2ab \cos C \end{aligned}$$

□

Since the first proof is simpler, we could view the second as a proof of the formula for sum of cosines, in reverse.

proof without words



Adding some words:

Draw a right triangle using one diagonal of a circle (any third point on the circle forms a right angle), then draw any other diagonal, in red.

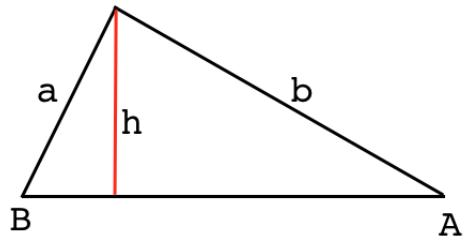
At the point of interest, the red diagonal is divided into $a + c$ and $a - c$. The base of the black right triangle is $2a \cos \theta$ so it is divided into b and $2a \cos \theta - b$.

We apply the theorem on chords of a circle from ([here](#)). The products of the chord segments are equal.

$$\begin{aligned}(a + c)(a - c) &= (2a \cos \theta - b)(b) \\ a^2 - c^2 &= 2ab \cos \theta - b^2 \\ c^2 &= a^2 + b^2 - 2ab \cos \theta\end{aligned}$$

Law of sines

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



$$\frac{h}{b} = \sin A$$

$$\frac{h}{a} = \sin B$$

Therefore

$$h = b \sin A = a \sin B$$

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

We could do the same construction and argument with either A or B , and the third angle, call it C . Therefore

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

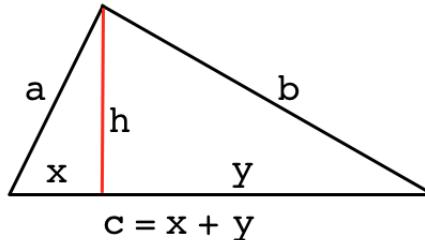
The sine of an angle, divided by the length of the side opposite, is a constant.

Chapter 12

Heron and Brahmagupta

Heron (or Hero) of Alexandria lived in the first century AD. He was primarily an engineer, but is also remembered for Heron's Formula, which can be used to compute the area of a triangle from the lengths of its sides. It is a simple formula that does not explicitly include the altitude h or the components of side c .

Heron's formula was later found to be a special case of a similar formula for quadrilaterals, discovered by Brahmagupta, which we'll study at the end of this chapter.



Let s be one-half the perimeter, called the semi-perimeter:

$$s = \frac{1}{2}(a + b + c)$$

$$2s = a + b + c$$

then Heron says that the area is

$$A = \sqrt{s \cdot (s - a) \cdot (s - b) \cdot (s - c)}$$

$$A^2 = s \cdot (s - a) \cdot (s - b) \cdot (s - c)$$

We first explore a sketch of a proof, which justifies why each term is present in the equation. To begin, note that the equation is symmetrical in a, b and c . This is expected, since there is no reason to distinguish among the sides.

Levi

Mark Levi has a short proof of Heron's formula, linked on this page:

<https://www.marklevimath.com/sinews>

This may be the original version, by two other authors

<https://sites.math.rutgers.edu/~zeilberg/mamarim/mamarimPDF/heron.pdf>

From the latter:

The area-squared is obviously a symmetric and homogeneous polynomial of degree 4 in a, b, c , divisible by $(a + b - c)(a + c - b)(b + c - a)$, since degenerate triangles have zero area.

Hence the area-squared divided by $(a + b - c)(a + c - b)(b + c - a)$ is a symmetric and homogeneous polynomial of degree 1 in a, b, c , and so is $(a + b + c)$ times some constant that must be 1 by considering, say, the 90, 45, 45 triangle.

Let's just play with the formula. Take what is under the square root above:

$$s \cdot (s - a) \cdot (s - b) \cdot (s - c)$$

Multiply each term by 2

$$\begin{aligned} & 2s \cdot (2s - 2a) \cdot (2s - 2b) \cdot (2s - 2c) \\ &= (a + b + c)(b + c - a)(a + c - b)(a + b - c) \end{aligned}$$

According to the formula above, $16A^2$, and hence the area itself, will be zero when

- o $a + b + c = 0$

that is, when the sum of all three sides is equal to zero. Since lengths are always positive, this means that $a = b = c = 0$, or

- o one of the other terms is zero, e.g. $a + b - c = 0$.

that is, when one side length is equal to the sum of the other two.

These are all "degenerate" triangles, where the shape has collapsed either to a point (the first case) or to a line segment.

The factor of 16 may be deduced from an example, e.g., an equilateral triangle with unit sides, altitude equal to $\sqrt{3}/2$ and area of $\sqrt{3}/4$.

Suppose we do not know the factor, so let it be k (rather than 16):

$$\begin{aligned} k \cdot \left(\frac{\sqrt{3}}{4}\right)^2 &= (a+b+c)(b+c-a)(a+c-b)(a+b-c) \\ &= 3 \cdot 1 \cdot 1 \cdot 1 = 3 \end{aligned}$$

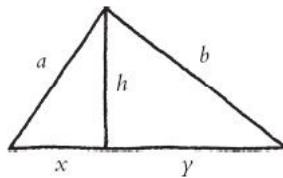
Clearly, $k = 4^2 = 16$.

The proof starts with the deduction that the area squared is "a polynomial of degree 4 in a, b, c ", and he works through why that is so. It makes sense, since area is itself the product of two lengths, each of which must be proportional somehow to the lengths of the sides.

Lockhart

Here is a seriously algebraic proof, from Lockhart. This is quite a long-winded approach, but there is a point!

Proof.



Side c is split into x and y . We can write three equations:

$$\begin{aligned} x^2 + h^2 &= a^2 \\ y^2 + h^2 &= b^2 \\ x + y &= c \end{aligned}$$

Our objective is an equation that contains only a , b and c . From the first two:

$$a^2 - b^2 = x^2 - y^2$$

and from the third:

$$y^2 = c^2 - 2xc + x^2$$

so

$$\begin{aligned} a^2 - b^2 &= x^2 - c^2 + 2xc - x^2 \\ &= 2xc - c^2 \end{aligned}$$

then

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

Finally a slight rearrangement:

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

This says that to find the point where c is divided into x and y , we move from the center $c/2$ a distance of $(a^2 - b^2)/2c$.

The corresponding equation for y is

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

which is easily checked by adding together the final two equations, obtaining $x+y = c$.

For the area, we will need h somehow. It is easier to use h^2 .

$$\begin{aligned} h^2 &= a^2 - x^2 \\ &= a^2 - \frac{(c^2 + a^2 - b^2)^2}{(2c)^2} \end{aligned}$$

The area squared is

$$\begin{aligned} A^2 &= \frac{1}{4}c^2h^2 \\ &= \frac{1}{4}c^2a^2 - \frac{1}{4}c^2\frac{(c^2 + a^2 - b^2)^2}{(2c)^2} \end{aligned}$$

Lockhart:

the algebraic form of this measurement is aesthetically unacceptable. First of all, it is not symmetrical; second, it's hideous. I simply refuse to believe that something as natural as the area of a triangle should depend on the sides in such an absurd way. It must be possible to rewrite this ridiculous expression...

Here's a start:

$$16A^2 = (2ac)^2 - (c^2 + a^2 - b^2)^2$$

This is much better. We immediately notice that it is a difference of squares. First

$$16A^2 = [2ac + (c^2 + a^2 - b^2)] [2ac - (c^2 + a^2 - b^2)]$$

But that has within it two squares, namely $(a+c)^2$ in the first term on the right-hand side, and $(a-c)^2$ in the second.

$$16A^2 = [(a+c)^2 - b^2] [b^2 - (a-c)^2]$$

A second difference of squares. Thus

$$16A^2 = (a+c+b)(a+c-b)(b+a-c)(b-a+c)$$

At this point, we recognize the semi-perimeter $2s = a + b + c$ and then we see that each of the other terms is s minus one of the sides. For example:

$$a + c - b = 2s - 2b$$

So we obtain

$$\begin{aligned} 16A^2 &= 2s \cdot (2s - 2a) \cdot (2s - 2b) \cdot (2s - 2c) \\ A^2 &= s \cdot (s - a)(s - b)(s - c) \end{aligned}$$

which you can of course write as the square root:

$$A = \sqrt{s \cdot (s - a)(s - b)(s - c)}$$

□

check

As a simple example, if we have a right triangle with sides 3,4,5, then the area is one-half of 3 times 4 = 6. The semi-perimeter is s

$$s = \frac{(3+4+5)}{2} = \frac{12}{2} = 6$$

We have

$$A = \sqrt{6(6-5)(6-4)(6-3)} = \sqrt{6(1)(2)(3)} = 6$$

from Twitter

I saw this variation on Twitter, but have lost the source.

In any triangle, twice the area is equal to the product of two sides times the sine of the angle between (since one side and the sine give the altitude). So

$$2A = ab \sin \gamma$$

where γ is the angle opposite the side with length c .

Now, the law of cosines ([ref](#)) says that, for example,

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

Rearranging

$$\cos \gamma = \frac{a^2 + b^2 - c^2}{2ab}$$

From Pythagoras

$$\begin{aligned} \sin \gamma &= \sqrt{1 - \cos^2 \gamma} \\ &= \sqrt{1 - \frac{(a^2 + b^2 - c^2)^2}{(2ab)^2}} \end{aligned}$$

The area is then:

$$2A = ab \sqrt{1 - \frac{(a^2 + b^2 - c^2)^2}{(2ab)^2}}$$

After squaring, we come to a similar double difference of squares as in the Lockhart proof (except that here we have terms like ab and there terms like ac).

So finally:

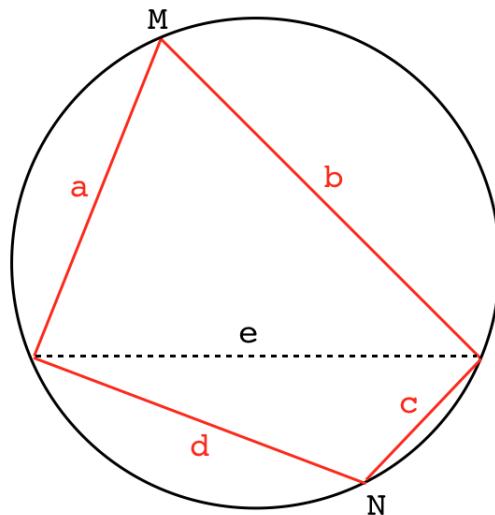
$$16A^2 = (a + b + c)(a + b - c)(c + a - b)(c - a + b)$$

□

Brahmagupta

Brahmagupta was an Indian mathematician who lived in the 7th century AD in a region of India called Bhinmal, which is in Rajasthan. He completed the square to obtain the quadratic equation, and did many other amazing things in trigonometry and arithmetic, as well as this example from geometry.

We consider a quadrilateral inscribed into a circle. This is a special case, where the fourth point fits into the same circle determined by any three of the points.



We will prove that the area of this quadrilateral is given by Brahmagupta's formula.

$$\begin{aligned} A &= \sqrt{(s-a) \cdot (s-b) \cdot (s-c) \cdot (s-d)} \\ A^2 &= (s-a) \cdot (s-b) \cdot (s-c) \cdot (s-d) \end{aligned}$$

Heron's formula is thus a special case where $d = 0$.

$$A = \sqrt{s \cdot (s - a) \cdot (s - b) \cdot (s - c)}$$

preliminary

We need two preliminary results. If M and N are supplementary angles, then

$$\sin M = \sin N, \quad \cos M = -\cos N$$

Supplementary angles have mirror image symmetry across the y -axis. This becomes obvious if you plot them.

Then, draw the line connecting the two opposing vertices which are not M and N . Using the law of cosines we can write two equal expressions for e^2 , namely:

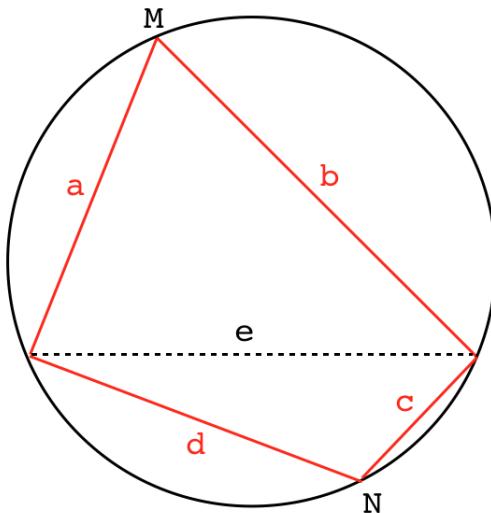
$$e^2 = a^2 + b^2 - 2ab \cos M$$

$$e^2 = c^2 + d^2 - 2cd \cos N = c^2 + d^2 + 2cd \cos M$$

Equating the two and grouping terms:

$$a^2 + b^2 - c^2 - d^2 = 2(ab + cd) \cos M$$

Look at the diagram again.



The triangle above the dotted line has area $(1/2) ab \sin M$ and similarly for the one below so the total area is

$$A_1 = \frac{1}{2}ab \sin M$$

$$A_2 = \frac{1}{2}cd \sin N = \frac{1}{2}cd \sin M$$

Adding, the total area is:

$$A = \frac{1}{2}(ab + cd) \sin M$$

$$4A = 2(ab + cd) \sin M$$

algebra

Square the two main equations so far:

$$(a^2 + b^2 - c^2 - d^2)^2 = [2(ab + cd)]^2 \cos^2 M$$

$$16A^2 = [2(ab + cd)]^2 \sin^2 M$$

and add

$$16A^2 + (a^2 + b^2 - c^2 - d^2)^2 = [2(ab + cd)]^2$$

Rearrange

$$16A^2 = [2(ab + cd)]^2 - (a^2 + b^2 - c^2 - d^2)^2$$

As before, we proceed to factor two differences of squares.

First:

$$16A^2 = [2(ab + cd) + (a^2 + b^2 - c^2 - d^2)] [2(ab + cd) - (a^2 + b^2 - c^2 - d^2)]$$

$$= [(a + b)^2 - (c - d)^2] [(c + d)^2 - (a - b)^2]$$

Second

$$= (a + b + (c - d))(a + b - (c - d)) (c + d + (a - b))(c + d - (a - b))]$$

$$= (a + b + c - d)(a + b - c + d)(c + d + a - b)(c + d - a + b)$$

If the semi-perimeter is s then

$$2s = a + b + c + d$$

So we have

$$\begin{aligned} 16A^2 &= (2s - 2d)(2s - 2c)(2s - 2b)(2s - 2a) \\ A^2 &= (s - d)(s - c)(s - b)(s - a) \end{aligned}$$

So lastly

$$\begin{aligned} A^2 &= (s - a)(s - b)(s - c)(s - d) \\ A &= \sqrt{(s - a)(s - b)(s - c)(s - d)} \end{aligned}$$

In comparing the two proofs, it's clear that Brahmagupta draws on the ideas of (i) using the semi-perimeter and (ii) difference of squares, which are in Heron's proof. The main new ideas are trigonometric: the law of cosines and the cancelation of $\sin^2 x + \cos^2 x$.

We can see this by rewriting the proof of Heron's formula in Brahmagupta's style.

But we don't have to! That's essentially what the proof from Twitter does, above. First, the law of cosines. Let α be the angle opposite side a :

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc \cos \alpha \\ (b^2 + c^2 - a^2)^2 &= (2bc)^2 \cos^2 \alpha \end{aligned}$$

And the area is

$$\begin{aligned} A &= (1/2)bc \sin \alpha \\ 4A &= 2bc \sin \alpha \\ 16A^2 &= (2bc)^2 \sin^2 \alpha \end{aligned}$$

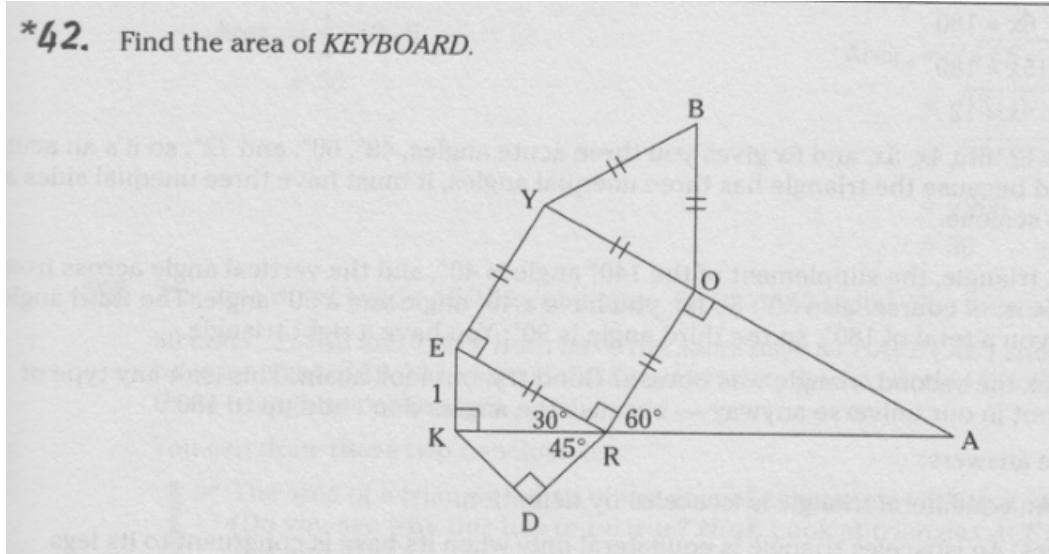
Adding

$$\begin{aligned} 16A^2 + (b^2 + c^2 - a^2)^2 &= (2bc)^2 \\ 16A^2 &= (2bc)^2 - (b^2 + c^2 - a^2)^2 \end{aligned}$$

The rest is exactly as before, it's just a matter of two differences of squares.

example

Here is a problem where we can use Heron's formula:



The smaller 30-60-90 right triangle has a side labeled 1.

Since $\sin 30^\circ = 1/2$, the side of the square has length 2 so the square has area 4.

Using Heron's formula, the equilateral triangle has area

$$A_{eq} = \sqrt{3} (1)^3 = \sqrt{3}$$

We get the base of the largest right triangle from the tangent of 60° .

$$\tan \pi/3 = \frac{\sin \pi/3}{\cos \pi/3} = \frac{1/2}{\sqrt{3}/2} = \frac{1}{\sqrt{3}}$$

so

$$\frac{2}{\text{base}} = \frac{1}{\sqrt{3}}$$

The base is $2\sqrt{3}$ and the area is then

$$A_{bigT} = \frac{1}{2} 2\sqrt{3} \cdot 2 = 2\sqrt{3}$$

Next is the small $\triangle EKR$. Its base is

$$\text{base} = 2 \cos \pi/6 = 2 \frac{\sqrt{3}}{2} = \sqrt{3}$$

$$A = \frac{1}{2} \sqrt{3} \cdot 1 = \frac{\sqrt{3}}{2}$$

Finally, the last triangle is isosceles. We know its diagonal is $\sqrt{3}$. Let the side be x , then

$$\frac{x}{\sqrt{3}} = \frac{1}{\sqrt{2}}$$

$$x = \frac{\sqrt{3}}{\sqrt{2}}$$

The area is

$$A = \frac{1}{2}x^2 = \frac{1}{2} \cdot \frac{3}{2} = \frac{3}{4}$$

The total is

$$4 + \sqrt{3} + 2\sqrt{3} + \frac{\sqrt{3}}{2} + \frac{3}{4}$$

which equals something.

Part IV

Two basic operations in calculus

Chapter 13

Simple slopes

To introduce the two fundamental ideas in calculus, consider two measuring devices used while driving a car. Most good drivers look fairly often at the speedometer, which measures speed or velocity, or how fast you're going.

On the other hand, if someone gives you directions like — go three and a half miles and then turn left (where the old gas station used to be), you need to be watching your odometer.



Distance divided by time is velocity. Velocity times time equals distance. We can think of speed and velocity as the same for now.

Velocity is the *rate of change* of distance with time, it has units like miles per hour or feet per second (15 mph is exactly 22 feet per second; $15 \cdot 5280 = 22 \cdot 3600$).

In calculus we say that

- velocity is the **derivative** of the distance with respect to time
- distance is the **integral** of the velocity with respect to time

We can speak of velocity at a particular time t , as in "our current velocity is 60 miles per hour." But the distance, the integral, must be evaluated between appropriate starting and stopping points for the time.

In our example, you must first look at your odometer *before* you start on that 3.5 mile drive, and subtract the initial from the final value.

time-dependence

Distance equals velocity times time.

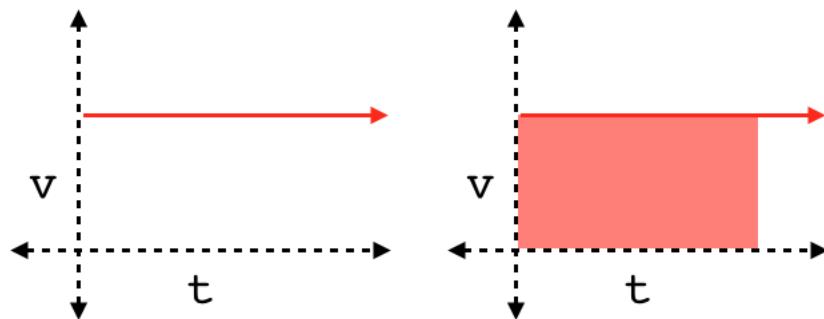
This is easy if the velocity is constant. Travel west on the interstate at exactly 60 miles per hour for 2 hours and your distance will be 120 miles from where you started (provided you don't start in Los Angeles).

It is standard to use s to refer to the distance traveled and v for velocity. If the velocity is constant then:

$$s = vt$$

According to the internet, s is from the Latin "spatium", for "space, room, or distance."

Suppose we plot velocity as a *function of time* with v on the y -axis and t on the x -axis.



Since the velocity is constant, the result is a straight horizontal line.

Furthermore, the distance traveled is the *area under the curve* (and above the x -axis)

which is the area of a rectangle with sides v and t and as we said

$$s = vt$$

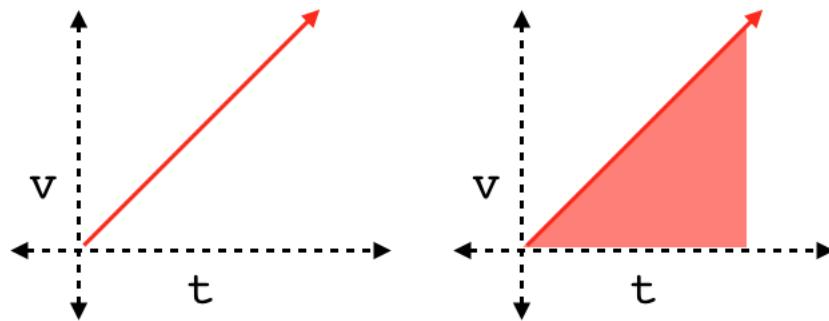
However, for most interesting problems the velocity is not constant.

Imagine maintaining pressure on the gas pedal in the car steadily so that, starting from a stop at zero time, after 1 second your velocity is 10 mph, after 2 seconds it is 20 mph, after 3 seconds, 30 mph. If we continue at the same rate of acceleration, we'll go from 0 to 60 mph in 6 seconds, which is quite a respectable time.

This example has constant acceleration. Here, we say that v is a constant function of time, and write

$$v = at$$

where a is the acceleration.



What about the distance? It turns out that the distance is once again the area under the curve.

If a is non-zero and constant, then v changes at a constant rate. Starting from time 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. We introduce subscripts to keep things straight. The initial velocity is v_i and the *final* velocity is v_f .

The distance traveled is the *average* velocity times the elapsed time. For smooth (constant) acceleration from zero to v , the average velocity is the average of the initial and final velocities:

$$v_{\text{avg}} = \frac{1}{2} (v_i + v_f) = \frac{1}{2} v_f$$

So the correct equation is:

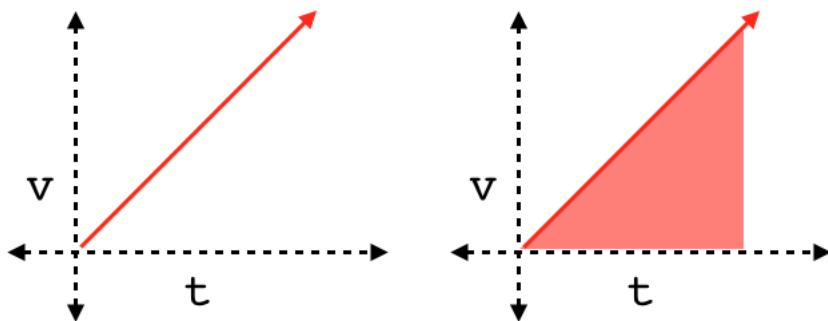
$$s = v_{\text{avg}} t = \frac{1}{2} v_f \cdot t$$

and since $v_f = at$

$$s = \frac{1}{2} a t^2$$

In this case, if we plot velocity as a function of time, we obtain a straight line that extends diagonally up with respect to the x -axis. The distance traveled is the area under the curve, below the line and above the x -axis.

The shape whose area is needed is a triangle. This also gives us the factor of $1/2$.



You probably know that if a mass m is dropped from a tall building like the Tower of Pisa, then the distance it has fallen goes like the square of the time. The equation is:

$$s = \frac{1}{2} g t^2$$

where g is the acceleration due to gravity.

Notice that this is the same equation as we just obtained.

The reason is that g is approximately constant near the surface of the earth, its value is about 10 in units of m/s^2 . A fall of four seconds is about 80 meters.

Galileo knew this formula (at least, he knew the t^2 part of it), which he obtained not from experiments at the Tower of Pisa, but by timing the descent of balls down an inclined plane.



initial position and velocity

If you want to be more complete and say that the starting point is not necessarily the origin of the coordinate system, add a constant s_0 to describe the initial distance from the origin and obtain:

$$s = vt + s_0$$

and similarly, a constant v_0 to describe the initial velocity as shown above.

The full equation of motion is

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

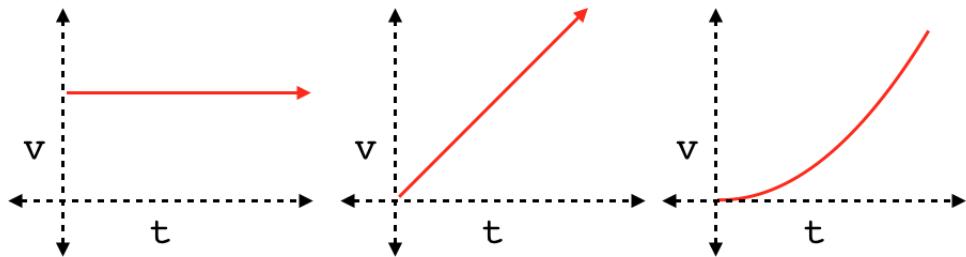
We'll say much more about this later.

power rule

We will introduce the theory of calculus more thoroughly in another book. For now, we just talk about a simple rule called the power rule.

In the previous section the variables were velocity and time (and distance). In general, velocity might be (i) constant, always the same independent of time, (ii) increasing linearly with time, (iii) dependent on the square of the time, or even (iv) some other power.

The first three of these are (respectively) the equations of: (i) a horizontal line, since v is constant, (ii) any other non-vertical line (v is proportional to t), and (iii), a parabola (v is proportional to t^2).



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.

- $y = c$
- $y = cx$
- $y = cx^2$

Suppose we are at some particular point on the curve, x .

We ask "what happens if we change x a little bit" and use the notation dx to refer to this little bit of x .

What happens to y ? y will usually change by a small amount. Call that amount dy .

case 0

We can call this case 0 because we can write it as

$$y = cx^0 = c$$

Of course, in this case

$$y = c$$

y does not actually depend on x at all. The change dy resulting from a change in x , dx , is zero. That is what the curve plotted above tells us (left panel).

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

The ratio dy/dx is the slope of the curve formed by plotting y against x . We call that slope the *derivative* of the function $f(x)$.

Divide both sides by dx and rewrite the above as:

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

This plot is a horizontal line with slope 0.

case 1

Here, y is a linear function of x , the change dy is the change dx multiplied by c :

$$y = cx, \quad dy = c \cdot dx$$

rearranging.

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

In analytical geometry, we calculate the slope of a line as $\Delta y/\Delta x$.

For a line, the slope is constant and so it doesn't matter which two points with coordinates $(x, y), (x', y')$ we choose for the calculation. The following is true for *any* two points on the line:

$$m = \frac{\Delta y}{\Delta x} = \frac{y - y'}{x - x'}$$

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

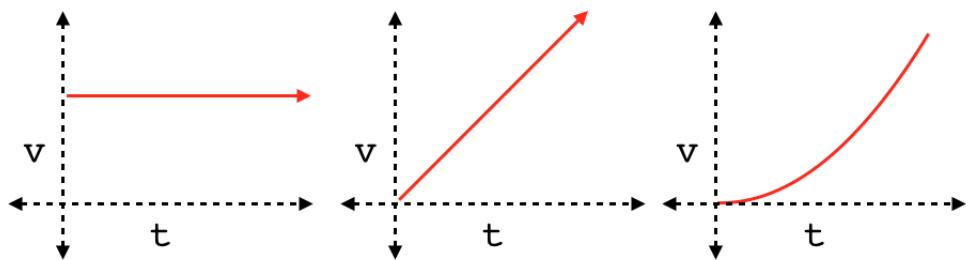
case 2

This case is different.

$$y = cx^2$$

We finally get to using some calculus.

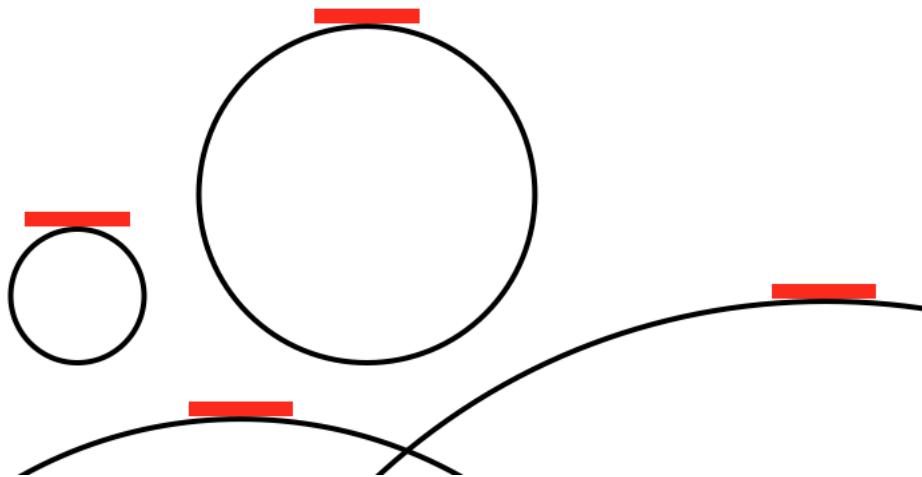
For a parabola, the slope of the curve at a point (the slope of the tangent to the curve $y = cx^2$) depends on the choice of x . The slope is steeper the further out you go in a positive direction on the x -axis (right panel).



It seems impossible to compute the slope of this curve in the standard way, by picking two points (x, y) and (x', y') and then calculating $\Delta y / \Delta x$, because the slope changes as we go out along the curve. You'll get a different answer for each different x .

key idea

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



the distance to the circle from the end of a line of fixed length decreases as we increase the size of the circle.

In calculus, we don't make the curves larger, we make the distance between x and x' smaller and smaller until it is so small, that the circle, or the parabola in the other figure, becomes flat. I have just magnified the figure so we could see it.

The error, the distance between the end of the red line above and the circle, gets smaller and smaller as a fraction of the line's length. Our approximation gets better

and better. Just zoom in until the line is a good enough approximation to the shape of the circle, if the curve doesn't look flat enough, zoom in some more.

As we are accelerating in the car, with constantly changing velocity, we can still have a unique velocity at a particular instant in time.

In mathematical language, for a very small change Δx in either direction from x , we get the same slope, *if* Δx is small enough.

If it's not, we can always make it smaller. That's the beauty of the real numbers.

As you accelerate from 0 to 60, there must be at least one moment in time when your velocity is 50.

Or, put still another way, when they built your house they didn't worry about the curvature of the earth.

If r is the radius of the earth in feet, and the house is $l = 50$ feet long, the drop due to curvature is $r - b$ where $b = \sqrt{r^2 - l^2}$

$$\begin{aligned} r &= 21120000 \\ b &= 21119999.9999408 \\ r-b &= 0.0000592 \end{aligned}$$

That is about 0.00006 feet over the length of a 50 foot house, about a thousand times less than 1/16 of an inch. It is nearly 6 orders of magnitude, one part in a million.

Since the changes in x and y are so small, we use the new nomenclature: dy and dx .

power rule again

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

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

$$v = \frac{ds}{dt} = at + v_0$$

$$\frac{dv}{dt} = a$$

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 .

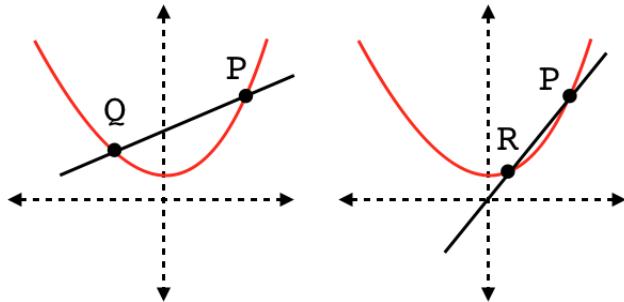
Well, you can. And I'll explain why as we go along.

Chapter 14

Difference quotient

In this chapter we look at the geometric interpretation of the derivative — which is the traditional way to begin calculus. The general approach was developed by Fermat.

Think for a minute about a curve such as the one shown in the figure, corresponding to some unspecified function $f(x)$, which looks like it's probably a parabola.



At an arbitrary point P on the curve, for some value of x , we plot $y = f(x)$. This is Descartes' genius idea. The point on the graph of $f(x)$ at x has coordinates $P = (x, f(x))$.

Now consider a point Q near P but also on the curve. For the x -coordinate of Q , a small change is made to x .

We might call that small amount Δx , but many authors use h , a simpler notation, and we will do so as well. The value of the function at $x + h$ is $f(x + h)$ and so Q

has coordinates $Q = (x + h, f(x + h))$.

In this example, h is negative, but that makes no difference. We drew it that way so it's easier to see how the approximation to the slope gets better as we go along.

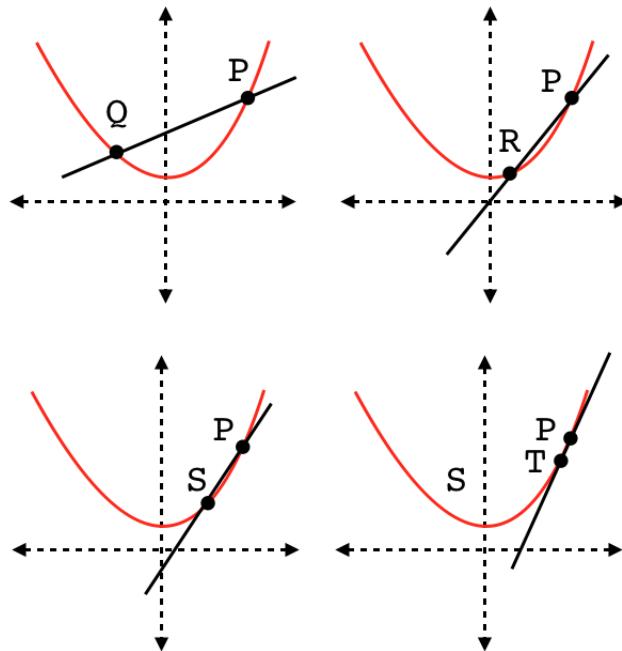
The slope of the (secant) line connecting Q and P is

$$\frac{\Delta y}{\Delta x} = \frac{f(x + h) - f(x)}{x + h - x} = \frac{f(x + h) - f(x)}{h}$$

This is a famous quantity, it's called the **difference quotient**.

The goal of differential calculus is to find the slope of the *tangent* to the curve at the point P . What we have is an expression for the slope of the secant line PQ , which is close but not quite the same thing.

To go from the secant to the tangent, we ask "what happens to this expression as h gets smaller and smaller and approaches zero." The second point where the secant meets the curve comes closer and closer to the first one.



In mathematical language, we say the slope of the tangent is equal to the limit of

the difference quotient as h tends to 0:

$$\lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

We'll say a bit more about limits in the next chapter, but for the moment you can think about

$$\lim_{h \rightarrow 0}$$

as meaning, "substitute $h = 0$ and see what happens to the expression of interest."

x squared

Let's try a couple of examples and look for a pattern.

$$f(x) = x^2$$

For this function, we write that the difference quotient is

$$\begin{aligned} & \frac{(x + h)^2 - x^2}{h} \\ &= \frac{x^2 + 2xh + h^2 - x^2}{h} \\ &= \frac{2xh + h^2}{h} \end{aligned}$$

Now divide by the denominator h

$$= 2x + h$$

Finally, to get the slope of the tangent, we evaluate the limit

$$\lim_{h \rightarrow 0} 2x + h = 2x$$

In evaluating the limit, we ask: what happens to this expression as h approaches 0. In this case, it cannot actually reach zero, because then our previous step of dividing by h would not be allowed. But we let h become really really small, and take advantage of the property of the limit which says that an expression can have a limit at c even if it can't be evaluated at c itself.

At every point on the curve $y = x^2$, the slope of the tangent line to the curve is $2x$. So the slope at $x = 0$ is 0, and the slope at $x = 2$ is 4, and so on.

This process of computing the difference quotient and then finding the limit as $h \rightarrow 0$ is called "taking the derivative." It produces an expression which is called the derivative of y with respect to x , in this case

$$\frac{dy}{dx} = 2x$$

and we can interpret this as the slope of the tangent to the curve of $f(x)$ at the point x .

Another useful shorthand uses the f from $f(x)$. We adopt the convention that the derivative of $f(x)$ can be written $f'(x)$.

$$f'(x) = 2x$$

To be even more succinct we might write y' for $f'(x)$.

If we repeat this exercise with a leading constant a (that is, for $f(x) = ax^2$), we find that every term in the numerator of the difference quotient will contain a , and the final result will be $2ax$. Constants just get carried through.

square root

Now look at the square root:

$$f(x) = \sqrt{x}, \quad (x \geq 0)$$

The difference quotient for this function is

$$\frac{\sqrt{x+h} - \sqrt{x}}{h}$$

Clean up the numerator by multiplying by the conjugate

$$\begin{aligned} & \frac{\sqrt{x+h} - \sqrt{x}}{h} \cdot \frac{\sqrt{x+h} + \sqrt{x}}{\sqrt{x+h} + \sqrt{x}} \\ &= \frac{x+h-x}{h(\sqrt{x+h} + \sqrt{x})} \end{aligned}$$

$$\begin{aligned}
&= \frac{h}{h(\sqrt{x+h} + \sqrt{x})} \\
&= \frac{1}{\sqrt{x+h} + \sqrt{x}}
\end{aligned}$$

We evaluate the limit

$$\frac{dy}{dx} = \lim_{h \rightarrow 0} \frac{1}{\sqrt{x+h} + \sqrt{x}} = \frac{1}{2\sqrt{x}}$$

inverse

Consider the inverse function

$$\begin{aligned}
f(x) &= 1/x, \quad (x \neq 0) \\
&\frac{\frac{1}{x+h} - \frac{1}{x}}{h}
\end{aligned}$$

Clean up the numerator

$$\begin{aligned}
&\frac{\frac{1}{x+h} - \frac{1}{x}}{h} \cdot \frac{(x)(x+h)}{(x)(x+h)} \\
&= \frac{x - (x+h)}{h(x)(x+h)} \\
&= \frac{-h}{h(x)(x+h)} \\
&= -\frac{1}{(x)(x+h)}
\end{aligned}$$

We evaluate the limit:

$$\begin{aligned}
&\lim_{h \rightarrow 0} -\frac{1}{(x)(x+h)} \\
&\frac{dy}{dx} = -\frac{1}{x^2}
\end{aligned}$$

There's a pattern here. We will use the notation $f'(x)$ to indicate the slope of the curve $f(x)$ at x

$$f(x) = x^2 \Rightarrow f'(x) = 2x$$

$$f(x) = \sqrt{x} = x^{1/2} \Rightarrow f'(x) = \frac{1}{2}x^{-1/2}$$

$$f(x) = \frac{1}{x} = x^{-1} \Rightarrow f'(x) = -\frac{1}{x^2} = -x^{-2}$$

The general formula is

$$f(x) = x^n \Rightarrow f'(x) = nx^{n-1}$$

This is easily proved (for integer n) using the binomial expansion for $(x + h)^n$ for integral n ($n \in 1, 2, \dots$). We need only the first three terms:

$$(x + h)^n = x^n + nx^{n-1}h + n\frac{(n-1)}{2}x^{n-2}h^2 + \dots$$

The key point is that the last term shown and all subsequent terms contain powers of h^2 or higher.

After division by h , for each of these terms there will remain one or more terms of h , and in the limit $\lim_{h \rightarrow 0}$ these become zero.

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{(x + h)^n - x^n}{h} \\ &= \lim_{h \rightarrow 0} \frac{x^n + nx^{n-1}h + n\frac{(n-1)}{2}x^{n-2}h^2 + \dots - x^n}{h} \\ &= \lim_{h \rightarrow 0} \frac{nx^{n-1}h + n\frac{(n-1)}{2}x^{n-2}h^2 + \dots}{h} \\ &= \lim_{h \rightarrow 0} nx^{n-1} + n\frac{(n-1)}{2}x^{n-2}h + \dots \\ &= nx^{n-1} \end{aligned}$$

Another question is what to do with a sum or difference of polynomials, such as

$$f(x) + g(x)$$

If you write out the difference quotient

$$\frac{f(x + h) - f(x) + g(x + h) - g(x)}{h}$$

everything can be exactly as before, just grouping all terms with $f(x)$ and those with $g(x)$ separately.

$$[f(x) + g(x)]' = f'(x) + g'(x)$$

We showed above by computing the difference quotient directly that

$$f(x) = \sqrt{x}$$

$$f'(x) = \frac{1}{2\sqrt{x}}$$

Here is another approach to the same problem. Consider

$$y = x^2$$

$$\frac{dy}{dx} = 2x$$

Solve for x as a function of y :

$$x = \sqrt{y}$$

We can do algebra with *differentials* (with some constraints):

$$\frac{dy}{dx} \frac{dx}{dy} = 1$$

$$2x \frac{dx}{dy} = 1$$

$$\frac{dx}{dy} = \frac{1}{2x} = \frac{1}{2\sqrt{y}}$$

In observing the inverse relationship, remember that x and y are related by the equation $y = x^2$. For example, when $x = 2$, $dy/dx = 2x = 4$.

Using the relationship $f(x)$, when $x = 2$, $y = 4$, and $dx/dy = 1/2\sqrt{y} = 1/2\sqrt{4} = 1/4$, which is indeed the inverse of 4.

In this last section, after solving for x as a function of y , y is the *independent* variable. We can switch back to our usual notation:

$$\frac{dy}{dx} = \frac{1}{2\sqrt{x}}$$

problem

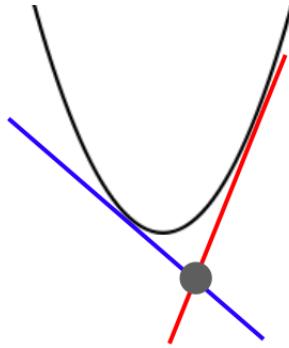
I found the following problems on the web. They are great practice and show what kinds of problems this approach of differentiation can solve. To prove:

Let $(a, f(a))$ and $(b, f(b))$ be two distinct points on the graph of a differentiable function f . Suppose that the tangent lines of f at these two points intersect, and call the point of intersection (c, d) . Verifying the following facts is elementary.

1. If $f(x) = x^2$, then $c = (a + b)/2$, the arithmetic mean of a and b .
2. If $f(x) = \sqrt{x}$, then $c = \sqrt{ab}$, the geometric mean of a and b .
3. If $f(x) = 1/x$, then $c = 2ab/(a + b)$, the harmonic mean of a and b .

1

Here is a diagram for the first one:



The claim is that the x -coordinate of the point will be half-way between the x -coordinates for the two points on the parabola. We have:

$$y = f(x) = x^2$$

$$y' = f'(x) = 2x$$

At $x = a$, the slope is $2a$ and the equation of a line through the point (a, a^2) is

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

At $x = b$, the equation is

$$y - b^2 = 2b(x - a)$$

To see where the lines cross, we set the y 's to be equal, and solve for x :

$$2a(x - a) + a^2 = 2b(x - b) + b^2$$

$$2ax - a^2 = 2bx - b^2$$

$$2x(a - b) = a^2 - b^2$$

$$= (a + b)(a - b)$$

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

2

We have:

$$y = f(x) = \sqrt{x}$$

$$y' = f'(x) = \frac{1}{2\sqrt{x}}$$

At $x = a$, the slope is $1/2\sqrt{a}$ and the equation of a line through the point (a, \sqrt{a}) is

$$y - \sqrt{a} = \frac{1}{2\sqrt{a}} (x - a)$$

At $x = b$, the equation is

$$y - \sqrt{b} = \frac{1}{2\sqrt{b}} (x - b)$$

We set the y 's to be equal

$$\frac{1}{2\sqrt{a}} (x - a) + \sqrt{a} = \frac{1}{2\sqrt{b}} (x - b) + \sqrt{b}$$

and solve for x . Multiply by $2\sqrt{a}\sqrt{b}$

$$(x - a)\sqrt{b} + 2a\sqrt{b} = (x - b)\sqrt{a} + 2b\sqrt{a}$$

Multiply through and cancel

$$x\sqrt{b} + a\sqrt{b} = x\sqrt{a} + b\sqrt{a}$$

$$\begin{aligned} x(\sqrt{b} - \sqrt{a}) &= b\sqrt{a} - a\sqrt{b} \\ &= \sqrt{a}\sqrt{b}(\sqrt{b} - \sqrt{a}) \end{aligned}$$

$$x = \sqrt{ab}$$

3

We have:

$$y = f(x) = \frac{1}{x}$$

$$y' = f'(x) = -\frac{1}{x^2}$$

At $x = a$, the slope is $-1/a^2$ and the equation of a line through the point $(a, 1/a)$ is

$$y - 1/a = -\frac{1}{a^2} (x - a)$$

At $x = b$, the equation is

$$y - 1/b = -\frac{1}{b^2} (x - b)$$

We set the y 's to be equal

$$-\frac{1}{a^2} (x - a) + 1/a = -\frac{1}{b^2} (x - b) + 1/b$$

and solve for x :

$$\left(\frac{1}{b^2} - \frac{1}{a^2}\right)x = 2\left(\frac{1}{b} - \frac{1}{a}\right)$$

$$\left(\frac{1}{b} + \frac{1}{a}\right)x = 2$$

$$(a + b)x = 2ab$$

$$x = \frac{2ab}{a + b}$$

Chapter 15

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

$$F(x) = \int f(x) \, dx = y$$

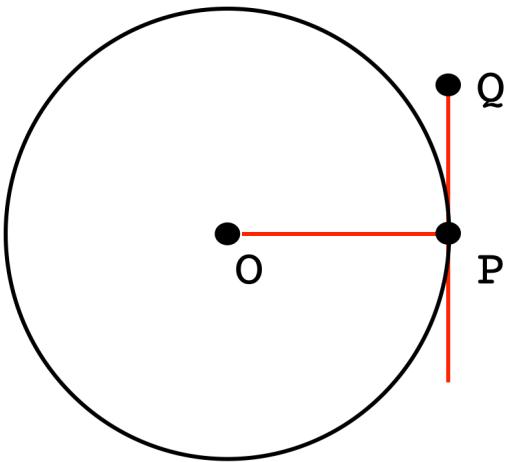
exactly when the derivative of $F(x)$ is $f(x)$:

$$\frac{dF}{dx} = F'(x) = f(x)$$

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 $1/2$.

$$= 2\pi \left[\frac{1}{2} r^2 \right] = \pi r^2$$

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

$$\begin{array}{c} r=R \\ | \\ r=0 \end{array}$$

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:

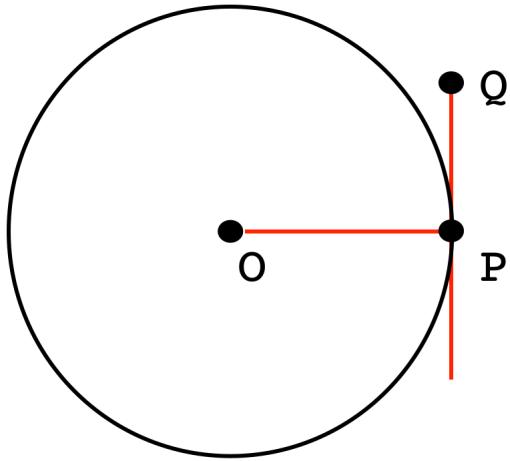
$$= \pi R^2 - \pi(0)^2 = \pi R^2$$

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

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

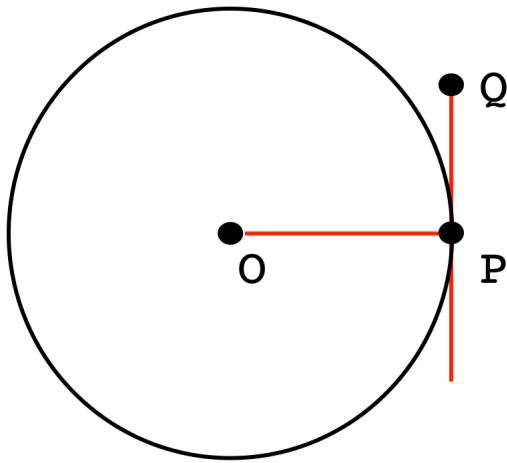
again we see that constants can come outside the integral

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

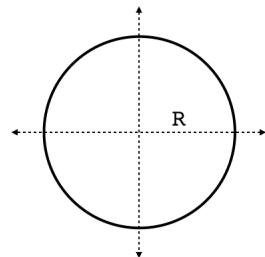
$$= \frac{1}{2} R^2 \theta \Big|_{\theta=0}^{\theta=2\pi}$$

$$= \pi R^2$$

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.

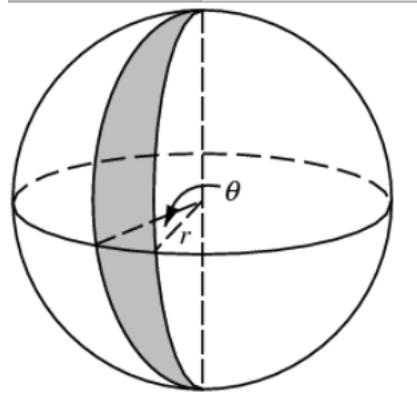
Volume of the sphere

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

$$\begin{aligned} dV &= 4\pi r^2 \, dr \\ V &= \int dV = \int_0^R 4\pi r^2 \, dr \\ &= 4\pi \left. \frac{1}{3}r^3 \right|_0^R = \frac{4}{3}\pi R^3 \end{aligned}$$

It's really as simple as that. Of course, you need to know the formula for the surface area to do it that way. Alternatively, if you know the volume of the sphere, taking the derivative is an easy way to get a formula for the surface area.

The image shows a "spherical lune", or segment of the surface of the sphere, as an aid to visualizing the whole surface.



We'll say a lot more about the volume of the sphere **later**.

technical note

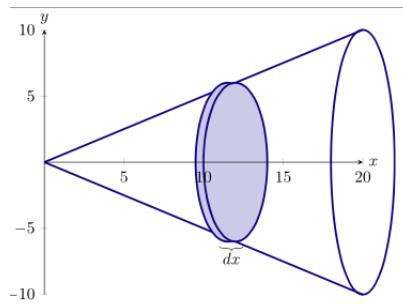
We should point out that this connection between volume and surface area is not true for *every* solid.

As an example, the surface area of a cube of side s is $6s^2$, which would have volume $2s^3$ if the relationship were always correct. In fact, there is something special about the *radial symmetry* of circles and spheres, and their lack of sharp corners and edges.

Here is one more example, to calculate the volume of a cone.

volume of a cone

We lay a cone along the x -axis with its vertex at the origin, opening to the right.



The cone is three-dimensional with the third axis (z) coming up out of the page. The intersection with the xy -plane is a triangle.

Can you see that in the xy -plane y is a linear function of x , i.e. $y = kx$ where k is a constant. The constant k is actually the ratio of the radius R to the height H . That is equal to $\Delta y/\Delta x$.

$$y = \frac{R}{H}x$$

If we slice the cone into thin sections perpendicular to the x -axis, each little piece is a circle with radius y and area πy^2 . For a thin enough slice, the volume is that area times the width of the slice:

$$dV = \pi y^2 dx$$

Finding the volume of an individual piece is the important part of the calculus argument.

Now we just substitute the value of y in terms of x

$$dV = \pi \left[\frac{R}{H} \right]^2 x^2 dx$$

add up all the little volumes by setting up the integral

$$V = \int dV = \int \pi \left[\frac{R}{H} \right]^2 x^2 dx$$

We apply the basic rule that constant terms can move "out from under" the integral sign:

$$= \pi \left[\frac{R}{H} \right]^2 \int x^2 dx$$

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 :

$$= \pi \left[\frac{R}{H} \right]^2 \int_0^H x^2 dx$$

and then just follow the rule for doing a problem like this: $\int x^2 = x^3/3$. So

$$= \pi \left[\frac{R}{H} \right]^2 \left[\frac{x^3}{3} \right] \Big|_0^H$$

$$= \frac{1}{3}\pi R^2 H$$

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

$$\begin{aligned} &= \pi \left[\frac{R}{H} \right]^2 \left[\frac{x^3}{3} \right] \Big|_{h_1}^{h_2} \\ &= \pi \left[\frac{R}{H} \right]^2 \left[\frac{h_2^3}{3} - \frac{h_1^3}{3} \right] \end{aligned}$$

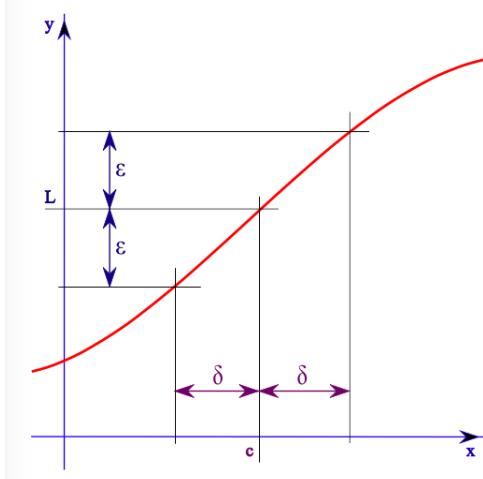
The geometers have given us an even more elegant formula ([here](#)).

Chapter 16

Limit concept

Limit concept

Consider the graph of a function $f(x)$. We might choose a power of x similar to $y = x^2$ or $y = x^3 - x$, which affirmatively has two properties that are of interest here: continuity and differentiability (we'll get to those ideas in a bit). Let's just say $y = f(x)$ is a "good" function. The functions we deal with in this book are all "good."



Focus on the neighborhood of a point on the x -axis, $x = c$.

By inspection of the graph, for points near c , the value of f at those points is not

too different from L .

(It is also true here that the value of $f(x)$ at c is equal to L . This matters for continuity but not for limits).

We would like to say that the *limit* of $f(x)$ as x approaches c is equal to L . The idea is that we can make $f(x)$ as close to L as we please, provided we choose x sufficiently close to c .

When the values successively attributed to a variable approach indefinitely to a fixed value, in a manner so as to end by differing from it by as little as one wishes, this last is called the limit of all the others. —Cauchy



Modern mathematicians don't like that word "approach", which conjures up movement and the involvement of time.

They also don't like reasoning from what they see in a graph, in part because no graph can show the whole function for the general case. To free ourselves from graphs and pictures, we will use an algebraic method from the formal apparatus of calculus.

There are two equivalent approaches, neighborhoods, and epsilon-delta formalism. Let's look at neighborhoods briefly.

neighborhoods

First, an *interval* between two real numbers a and b ($a < b$) contains every real number $a < x < b$.

$$(a, b) = x \mid a < x < b$$

The " | " means x "such that" the condition $a < x < b$ holds.

A *closed* interval $[a, b]$ includes the endpoints, $a \leq x \leq b$, while an *open* interval (a, b) excludes them. Half-open intervals like $[a, b)$ may be defined, and an interval with $\pm\infty$ as an endpoint is always open on that end, for example: $[a, \infty)$, because infinity *is not a number*.

Any open interval with a point p as its midpoint is called a *neighborhood* of p . Let r be the distance from p to the boundary of a particular neighborhood; r may be large or very very small. We denote a neighborhood of p as $N(p)$. $N(p)$ consists of all those values of x such that

$$|x - p| < r$$

which we would write more formally as

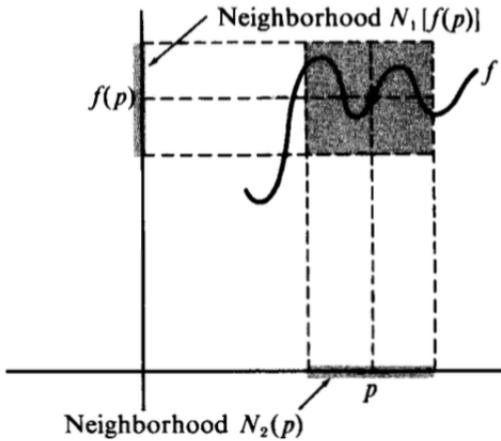
$$N(p) = x \mid |x - p| < r$$

To say that the limit $f(x) \rightarrow L$ exists, we mean that for every neighborhood $N_1(L)$, no matter how small, there exists some neighborhood $N_2(p)$ such that $f(x)$ is contained within $N_1(L)$, written as

$$f(x) \in N_1(L)$$

whenever $x \in N_2(p)$.

If $N_1(L)$ is very small, then $N_2(p)$ may need to be very small as well, to guarantee that $f(x)$ is contained within N_1 . Here is an example where this condition is satisfied.



The idea of a neighborhood is a nice abstraction to hide the apparatus of modern calculus, which we save for the Addendum.

An important fact about limits has to do with the case where $x = p$. It is *not* necessary that $f(p) = L$. This relaxed condition is in fact crucial for calculus.

example 1

Limits can be easy or hard, depending on the problem. Here is one found in the previous chapter on difference quotients:

$$\lim_{h \rightarrow 0} \frac{2xh + h^2}{h}$$

When you see something like this, what you are supposed to do is reason about what happens as the variable h approaches 0 (gets smaller and smaller). The first step in that is to figure out what would happen if h actually would become zero.

Here, each term has a limit of 0 when h is zero, so we will have 0/0. The zero on the bottom is trouble, it means that the expression becomes undefined.

However, suppose we first cancel h on top and bottom to obtain

$$\lim_{h \rightarrow 0} \frac{2x + h}{1} = \lim_{h \rightarrow 0} 2x + h$$

Now, the answer is just $2x$. This is valid as long as h approaches zero but is never actually equal to it.

Recall that we can have a limit for $f(x)$ as x approaches c , even if $f(c)$ does not exist.

example 2

Here is another important expression. What is the value of $f(x)$ as h approaches zero?

$$\cos h < f(x) < \frac{1}{\cos h}$$

Since $\cos 0 = 1$, the two outside terms both approach 1 in the limit as h approaches zero. Since $f(x)$ lies between them, it must also approach 1.

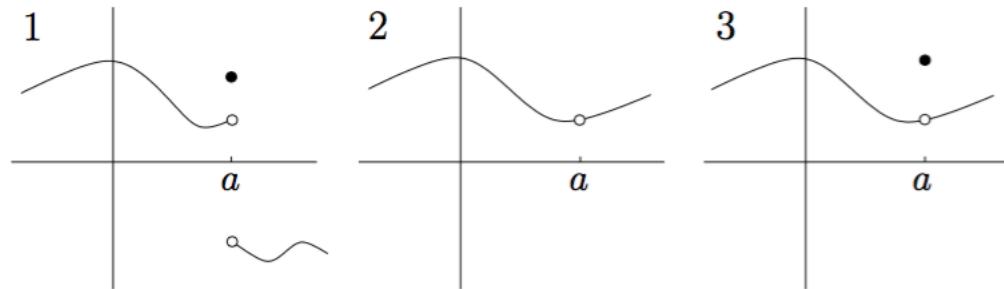
This is called the *squeeze theorem*.

The magical thing is that this is true even if, when $h = 0$, $x = 0/0$. We'll see this when we look at calculus of sine and cosine.

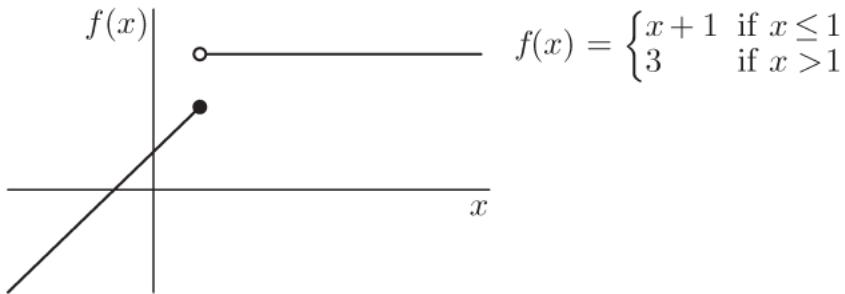
Continuity

Continuity has an intuitive definition: as Euler said, if we can graph a function *without lifting our pencil from the paper*, then the function is continuous.

Here are some graphs showing examples of how continuity can fail.



A filled circle means that the function yields that y -value for the corresponding x -value of the point, while an open circle means it does not. The function may yield some other value, or simply be undefined.



For a function to be continuous at a point $x = c$, we imagine that if we vary x in neighborhood of c , then $f(x)$ should not change in value by too much.

Again, we will call that value L , the limit of $f(x)$ as $x \rightarrow c$. For L to exist we require that the two one-sided limits be equal. If we approach c from the high side ($x > c$) or the low side ($x < c$), the limit must be the same.

Very important: continuity requires, in addition, that $f(c)$ be equal to L .

Differentiability

For a function to be differentiable, we require that the limit

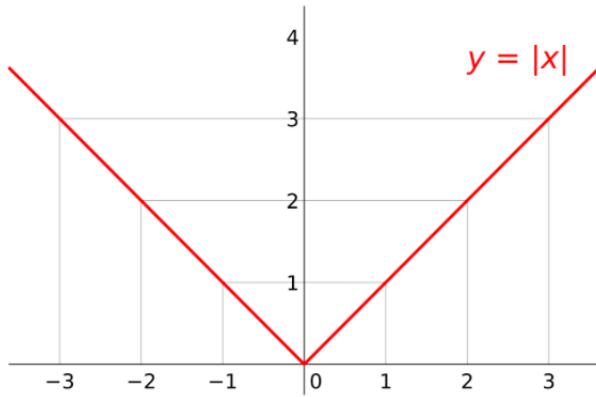
$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

exists. An example of a function that is continuous but not differentiable at a particular point is the absolute value function.

example: absolute value

An algebraic definition of the absolute value function is piecewise:

$$|x| = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$$



The function $f(x) = |x|$ is continuous at $x = 0$ because the two one-sided limits exist and are equal to each other. They are also equal to $f(0) = 0$.

However, there is no defined slope at $x = 0$. The difference quotient gives different results for positive Δx (positive slope) than for negative Δx (negative slope).

Without getting too technical

Note that the graph of the absolute value function is "all in one piece", but has a "sharp point" at the origin. We will not attempt to make these descriptions precise, other than to say that the fact that the graph comes "all in one piece" is a feature of continuity, and that graphs of differentiable functions are "smooth" in that they do not have "sharp points." The unambiguous and demonstrably true statement here is that the absolute value function is continuous at 0 but is not differentiable at 0.

<https://oregonstate.edu/instruct/mth251/cq/Stage5/Lesson/diffVsCont.html>

practical limits

- Plug in the value and see what happens. No problem here:

$$\lim_{x \rightarrow 2} \frac{x+1}{x^2+3} = \frac{3}{7}$$

- Division by zero isn't allowed. But we can factor:

$$\lim_{x \rightarrow 3} \frac{x^2 - 9}{x - 3} = \lim_{x \rightarrow 3} \frac{(x + 3)(x - 3)}{x - 3} = \lim_{x \rightarrow 3} x + 3 = 6$$

- Limit at infinity. Convert to a limit at zero:

$$\lim_{x \rightarrow \infty} \frac{x^2 + 3}{3x^2 + x + 1} = \lim_{1/x \rightarrow 0} \frac{1 + 3/x^2}{3 + 1/x + 1/x^2} = \frac{1}{3}$$

Part V

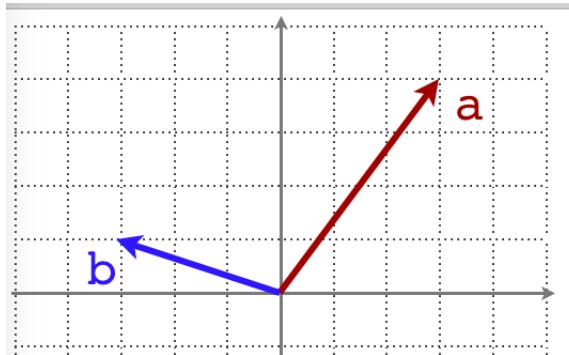
Vectors

Chapter 17

Vector dot product

In this chapter, we look at a few useful properties and operations of vectors in two- and three-dimensional space. I assume that you have already encountered vectors before, so this is not totally new.

From a geometrical point of view, a vector is a mathematical object that has both magnitude and direction. For example, in the standard 2D-coordinate system, the (maroon) vector $\langle 3, 4 \rangle$ goes out from the origin three units in the x -direction and four units in the y -direction.



Vectors are written in bold type:

$$\mathbf{a} = \langle 3, 4 \rangle$$

$$\mathbf{b} = \langle -3, 1 \rangle$$

A vector has one property of a line, slope, but the fixed magnitude means that a

vector does not extend to infinity as a line does. The squared length of a vector can be computed as the sum of the squares of its components, according to Pythagoras.

$$(\text{length } \mathbf{a})^2 = |\mathbf{a}|^2 = 3^2 + 4^2$$

By convention, we allow vectors to move about in space. We mean that two vectors of the same length, and pointing in the same direction are considered to be the same object, regardless of where they are located in space. (Some physics problems don't allow this, but in math it's the usual case).

So if we have the vector $\mathbf{v} = \langle 1, 1 \rangle$ starting at the origin $(0, 0)$ and ending at the point $(1, 1)$, and compare it to a second vector \mathbf{u} that starts from $(2, 0)$ and ends at $(3, 1)$, those are considered to be the same vector.

As you might guess, the vector that connects two points (x_1, y_1) and (x_2, y_2) is

$$\mathbf{p} = \langle x_2 - x_1, y_2 - y_1 \rangle$$

If we do the subtraction in reverse we have

$$\mathbf{q} = \langle x_1 - x_2, y_1 - y_2 \rangle$$

$$\mathbf{p} = -\mathbf{q}$$

Vectors add by adding their components:

$$\mathbf{a} = \langle 3, 4 \rangle$$

$$\mathbf{b} = \langle -3, 1 \rangle$$

$$\mathbf{a} + \mathbf{b} = \langle 0, 5 \rangle$$

Subtraction works the same way.

From a linear algebra point of view, a vector is simply an ordered collection of numbers

$$\mathbf{u} = \langle u_1, u_2, \dots, u_n \rangle$$

where n could be very large, even infinite.

However, a lot of work is done in two or three dimensions (officially \mathbb{R}^2 and \mathbb{R}^3), and the principles developed there carry over nicely into n -dimensional space. So let's start by thinking about a two-dimensional vector

$$\mathbf{u} = \langle u_1, u_2 \rangle$$

As I've said, the vector \mathbf{u} can be thought of as an arrow that goes from the origin to the point (u_1, u_2) . It has both length and direction, with the length given by

$$|\mathbf{u}| = \sqrt{u_1^2 + u_2^2}$$

and its direction is

$$\frac{u_2}{u_1} = \tan \theta, \quad \theta = \tan^{-1} \frac{u_2}{u_1}$$

where θ is the angle the vector makes (rotating counter-clockwise) from the positive x-axis.

Any vector can be converted into a *unit vector*, a vector of length one, by dividing by its length. For example if $\mathbf{u} = \langle 1, 2 \rangle$ then

$$\hat{\mathbf{u}} = \frac{1}{|\mathbf{u}|} \mathbf{u} = \left\langle \frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}} \right\rangle$$

$\hat{\mathbf{u}}$ is a unit vector pointing in the same direction as \mathbf{u} .

The line through the origin with slope $m = u_2/u_1$ and equation

$$y = mx$$

can be thought of as the extension of vector \mathbf{u} obtained by multiplying some t times \mathbf{u} for all $t \in \mathbb{R}$. We have stretched the vector to infinity, and beyond!

The standard unit vectors point in the direction of the x , y and z axes.

$$\hat{\mathbf{i}} = \langle 1, 0, 0 \rangle$$

$$\hat{\mathbf{j}} = \langle 0, 1, 0 \rangle$$

$$\hat{\mathbf{k}} = \langle 0, 0, 1 \rangle$$

We can write the vector using these unit vectors as

$$\mathbf{a} = \langle 3, 4 \rangle = 3 \cdot \hat{\mathbf{i}} + 4 \cdot \hat{\mathbf{j}}$$

Dot product

We now introduce a procedure for multiplying two vectors, the *dot product*, and derive the relationship between the dot product of two vectors and the angle between them. Suppose we have two vectors

$$\begin{aligned}\mathbf{a} &= \langle a_1, a_2 \rangle \\ \mathbf{b} &= \langle b_1, b_2 \rangle\end{aligned}$$

Geometrically, we might think of these as being one vector extending from the origin in the x, y -plane to the point (a_1, a_2) , and the other vector extending from the origin to (b_1, b_2) . The dot product is defined as

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2$$

We can extend this to a pair of vectors in n -dimensional space

$$\begin{aligned}\mathbf{a} &= \langle a_1, a_2, \dots, a_n \rangle \\ \mathbf{b} &= \langle b_1, b_2, \dots, b_n \rangle \\ \mathbf{a} \cdot \mathbf{b} &= a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=0}^n a_i b_i\end{aligned}$$

The two vectors being multiplied (whose dot product is computed) must have the same dimension, the same n . Also, the result of the multiplication—the dot product—is a number. This is in contrast to another form of vector multiplication (the cross-product) which yields a vector as the result.

notation

The dot (\cdot) in the dot product may also be used to set apart two multiplicands in scalar multiplication, to increase clarity. So, you ask, how can we tell what is meant? Well, consider

$$\begin{aligned}v \cdot \frac{1}{v} \\ \mathbf{a} \cdot \mathbf{b}\end{aligned}$$

It's a dot product if the two objects are vectors, otherwise it's multiplication.

Some properties

The dot product obeys the usual rules: it is associative, commutative and distributive.

The commutative property of the dot product:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$$

follows from the same property for multiplication of real numbers, since

$$\begin{aligned}\mathbf{a} \cdot \mathbf{b} &= \sum_n a_n b_n \\ &= \sum_n b_n a_n = \mathbf{b} \cdot \mathbf{a}\end{aligned}$$

For the distributive property, suppose

$$\mathbf{b} = \mathbf{c} + \mathbf{d}$$

Then

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot (\mathbf{c} + \mathbf{d}) = \mathbf{a} \cdot \mathbf{c} + \mathbf{a} \cdot \mathbf{d}$$

You can easily verify this by computing each term of the respective products.

$$\begin{aligned}\mathbf{b} &= \langle b_1, b_2 \rangle = \mathbf{c} + \mathbf{d} = \langle c_1 + d_1, c_2 + d_2 \rangle \\ \mathbf{a} \cdot \mathbf{b} &= a_1(c_1 + d_1) + a_2(c_2 + d_2) \\ &= a_1c_1 + a_1d_1 + a_2c_2 + a_2d_2 \\ &= a_1c_1 + a_2c_2 + a_1d_1 + a_2d_2 \\ &= \mathbf{a} \cdot \mathbf{c} + \mathbf{a} \cdot \mathbf{d}\end{aligned}$$

Another example that we will need below is

$$(\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) = \mathbf{a} \cdot \mathbf{a} - \mathbf{a} \cdot \mathbf{b} - \mathbf{b} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b}$$

by the commutative property

$$= \mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} - 2 \mathbf{a} \cdot \mathbf{b}$$

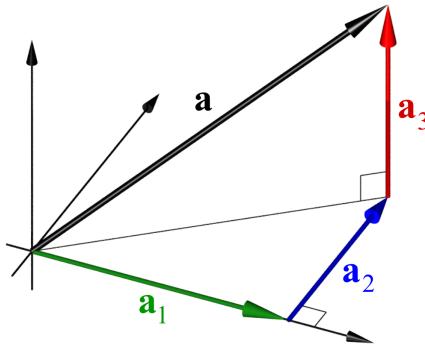
Length of a vector

As we said, the length of a vector $\mathbf{a} = \langle a_1, a_2 \rangle$, designated $|\mathbf{a}|$, is computed by a straightforward application of the Pythagorean Theorem:

$$|\mathbf{a}|^2 = a_1^2 + a_2^2$$

We leave the result as the square for simplicity.

This is easily extended to more dimensions by sequential application of the same method.



In \mathbb{R}^3 :

$$|\mathbf{a}|^2 = a_1^2 + a_2^2 + a_3^2$$

In \mathbb{R}^n :

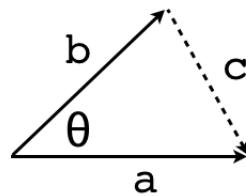
$$|\mathbf{a}|^2 = a_1^2 + a_2^2 + \cdots + a_n^2$$

Notice that

$$|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$$

Relation to θ

Now we are ready for the main idea. Suppose we draw two vectors \mathbf{a} and \mathbf{b} in \mathbb{R}^2 with their tails at the same point. Designate the angle between them as θ and the vector representing the side opposite as \mathbf{c} .



The orientation of \mathbf{c} doesn't matter for the argument that follows. As shown

$$\mathbf{b} + \mathbf{c} = \mathbf{a}$$

$$\mathbf{c} = \mathbf{a} - \mathbf{b}$$

Compute the dot product of \mathbf{c} with itself

$$\mathbf{c} \cdot \mathbf{c} = (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b})$$

Recalling the result from above, this is

$$\mathbf{c} \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} - 2 \mathbf{a} \cdot \mathbf{b}$$

Since

$$|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$$

and so on, we have that

$$\mathbf{c} \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} - 2 \mathbf{a} \cdot \mathbf{b}$$

$$|\mathbf{c}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2 \mathbf{a} \cdot \mathbf{b}$$

Does this remind you of the [law of cosines](#)?

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

Comparing the two equations, we see that

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta$$

This relationship is extremely useful because it allows us to compute the cosine of the included angle via the dot product.

Even more important, two vectors which are perpendicular will have $\cos \theta = 0$, so their dot product is zero. Two vectors pointed in the same direction have $\cos \theta = 1$ so it's just the product of the magnitudes.

This result extends to vectors in \mathbb{R}^n . Proof: choose a coordinate system where the two vectors lie in the same plane. Then apply the standard method.

For example, suppose I have the vector

$$\mathbf{u} = \langle p, q \rangle$$

Find a vector \mathbf{v} perpendicular to \mathbf{u} .

$$\mathbf{v} = \langle q, -p \rangle$$

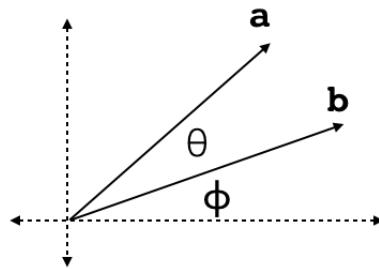
\mathbf{v} is perpendicular to \mathbf{u} because

$$\mathbf{u} \cdot \mathbf{v} = pq + q(-p) = 0$$

How to find a vector in \mathbb{R}^5 perpendicular to $\langle 1, 1, 1, 1, 0 \rangle$? Any vector of the form $\langle 0, 0, 0, 0, k \rangle$ will do, where k is some real number.

Alternate derivation

Here is another approach which doesn't depend on knowing the law of cosines, but uses the addition rule for cosine instead.



Vector \mathbf{a} forms an angle θ with vector \mathbf{b} . \mathbf{b} forms an angle ϕ with the x -axis, so the angle between \mathbf{a} and the x -axis is $\theta + \phi$.

Find the dot product using components. If $a = |\mathbf{a}|$ and $b = |\mathbf{b}|$ then

$$a_x = a \cos(\theta + \phi)$$

$$b_x = b \cos \phi$$

$$a_y = a \sin(\theta + \phi)$$

$$b_y = b \sin \phi$$

So

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= a_x b_x + a_y b_y \\ &= ab [\cos(\theta + \phi) \cos \phi + \sin(\theta + \phi) \sin \phi] \end{aligned}$$

Using the rule

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

the part in parentheses is

$$\begin{aligned} & \cos(\theta + \phi) \cos \phi + \sin(\theta + \phi) \sin \phi \\ &= \cos(\theta + \phi - \phi) = \cos \theta \end{aligned}$$

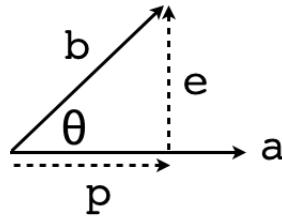
Another important property is that the value of the dot product is *independent* of the coordinate system chosen, because rotation or translation cannot change the lengths of the vectors nor the angle between them.

Projection

If $|\mathbf{a}| = 1$ we say that \mathbf{a} is a *unit vector*. In that case

$$\mathbf{b} \cdot \mathbf{a} = |\mathbf{b}| \cos \theta$$

Looking at the figure, $|\mathbf{b}| \cos \theta$ is the length of the *projection* of \mathbf{b} on \mathbf{a} . (Recall that the dot product is a scalar—a number—and not a vector).



The result, $\mathbf{b} \cdot \mathbf{a} = |\mathbf{b}| \cos \theta$, is the length of the part of \mathbf{b} that extends in the same direction as \mathbf{a} . The corresponding vector is

$$\mathbf{p} = (\mathbf{b} \cdot \mathbf{a}) \mathbf{a}$$

The other component of \mathbf{b} is the part that is perpendicular to \mathbf{p}

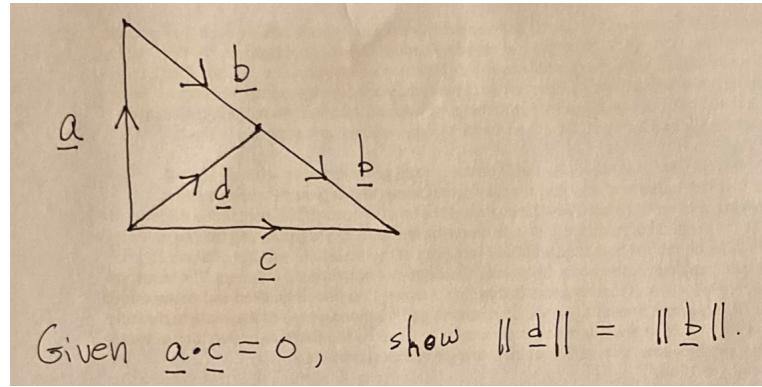
$$\mathbf{p} + \mathbf{e} = \mathbf{b}$$

We compute \mathbf{e} as the difference $\mathbf{b} - \mathbf{p}$. \mathbf{e} is the part of \mathbf{b} that is perpendicular to the projection. As a final note, the formula given here is a simplification for the situation in which \mathbf{a} is a unit vector. If not, the complete formula is:

$$\mathbf{p} = \frac{\mathbf{b} \cdot \mathbf{a}}{\mathbf{a} \cdot \mathbf{a}} \mathbf{a}$$

Vectors allow simple proofs for some geometric theorems such as Ceva's theorem and the law of cosines.

from Strogatz



This is the *midpoint* theorem for a right triangle.

$$\mathbf{a} = \mathbf{d} - \mathbf{b}$$

$$\mathbf{c} = \mathbf{d} + \mathbf{b}$$

Given

$$\mathbf{a} \cdot \mathbf{c} = 0$$

$$= (\mathbf{d} - \mathbf{b}) \cdot (\mathbf{d} + \mathbf{b})$$

$$= d^2 + b^2$$

$$d = b$$

from Strang

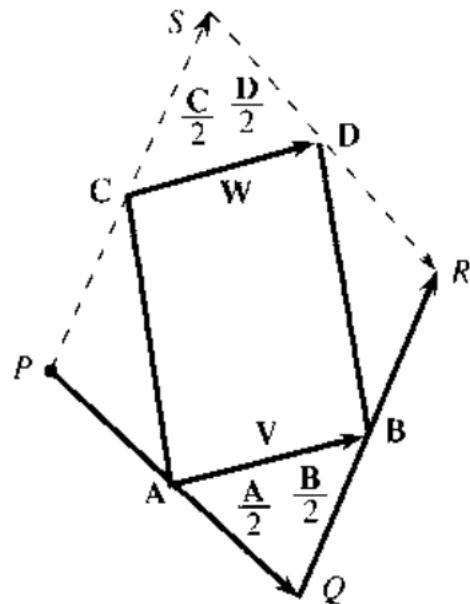


Fig. 11.4 Four midpoints

Consider *any* four-sided figure in space, such as $PQRS$ in the figure. (Note: $|\mathbf{A}| \neq |\mathbf{B}|$, and so on, and S is not co-planar with P, Q, R .

I claim that the midpoints of the sides form a parallelogram $ABCD$.

We will prove that $\mathbf{V} = \mathbf{W}$.

The figure makes it almost obvious.

$$\mathbf{V} = \frac{\mathbf{A}}{2} + \frac{\mathbf{B}}{2}$$

$$\mathbf{W} = \frac{\mathbf{C}}{2} + \frac{\mathbf{D}}{2}$$

The segment from P to R can be covered in two ways

$$\mathbf{A} + \mathbf{B} = \mathbf{C} + \mathbf{D}$$

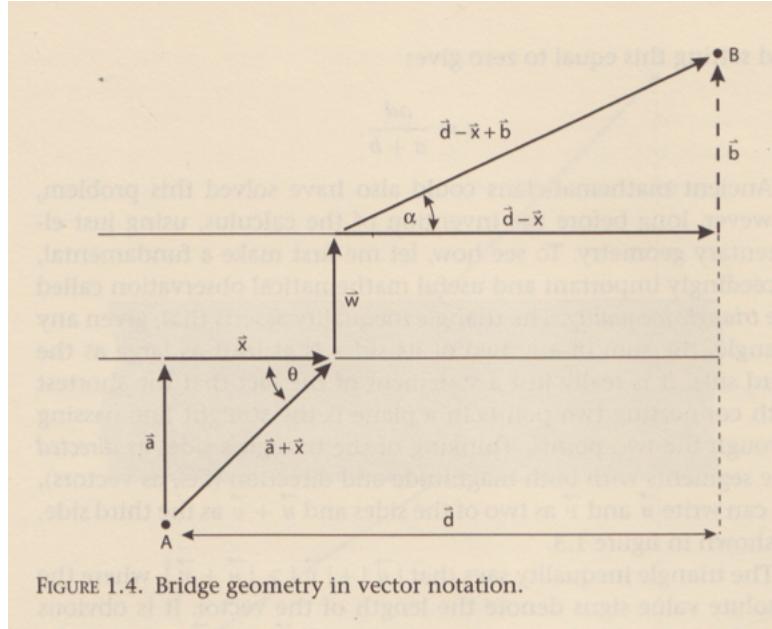
Divide both sides by 2 and obtain

$$\frac{\mathbf{A}}{2} + \frac{\mathbf{B}}{2} = \frac{\mathbf{C}}{2} + \frac{\mathbf{D}}{2}$$

$$\mathbf{V} = \mathbf{W}$$

□

from Nahin



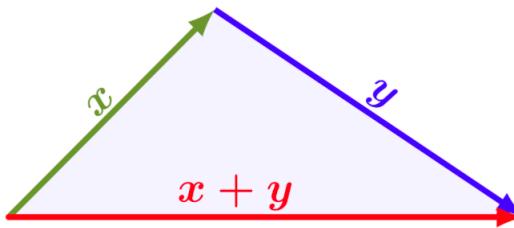
Two towns are on opposite sides of a river at points A and B . It is desired to choose the site of a bridge so as to minimize the distance between the two towns when traveling over the bridge. The problem can be set up algebraically and solved by differential calculus. However, the vector approach is more fun, and allows us to introduce the important *triangle inequality*.

Vectors are shown in the figure: \mathbf{a} is the perpendicular distance from A to the river, and similarly for \mathbf{b} . \mathbf{x} determines the placement of the bridge. If the horizontal distance between A and B is \mathbf{d} , then $\mathbf{d} - \mathbf{x}$ is the horizontal distance between B and the bridge. The distance across the bridge is \mathbf{w} , which cannot be changed. Its length will just be added onto our shortest path.

We want to choose \mathbf{x} so that the path from A to B is the shortest. The path from A to the bridge is $\mathbf{a} + \mathbf{x}$, that from the bridge to B is $\mathbf{b} + \mathbf{d} - \mathbf{x}$ so all together we have (taking the lengths of the vectors)

$$L = |\mathbf{a} + \mathbf{x}| + |\mathbf{b} + \mathbf{d} - \mathbf{x}|$$

The triangle inequality says that the lengths of two sides of a triangle add to be larger than or equal to the length of the third side.



$$|\mathbf{x}| + |\mathbf{y}| \geq |\mathbf{x} + \mathbf{y}|$$

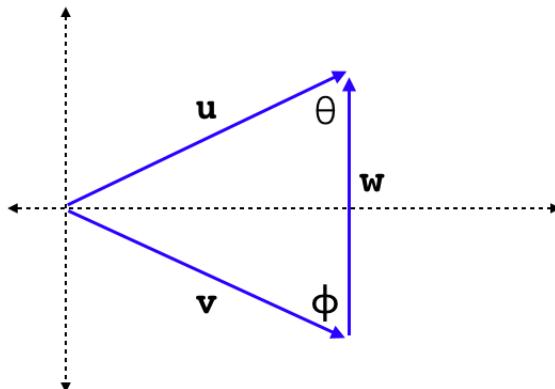
The rule is that the minimal value for the sum $|\mathbf{x}| + |\mathbf{y}|$ occurs when they point in the same direction.

In our problem, the minimum length occurs when $\mathbf{a} + \mathbf{x}$ and $\mathbf{b} + \mathbf{d} - \mathbf{x}$ point in the same direction. In other words, when $\theta = \alpha$.

Then, by similar triangles,

$$\begin{aligned}\frac{x}{a} &= \frac{d-x}{b} \\ bx &= ad - ax \\ x &= \frac{ad}{a+b}\end{aligned}$$

from Euclid



We are given a triangle with two sides the same length (isosceles). Without loss of generality, draw the triangle with its vertex at the origin and the midpoint of the third side on the x -axis.

To prove: $\theta = \phi$.

Let

$$\mathbf{u} = \langle a, b \rangle$$

$$\mathbf{v} = \langle a, -b \rangle$$

$$\mathbf{w} = \langle 0, 2b \rangle$$

We compute the dot products so that the angle between the vectors is acute and the dot product is > 0 .

$$\begin{aligned}\mathbf{u} \cdot \mathbf{w} &= 2b^2 \\ &= |\mathbf{u}| |\mathbf{w}| \cos \theta = \sqrt{a^2 + b^2} \cdot 2b \cos \theta \\ \cos \theta &= \frac{b}{\sqrt{a^2 + b^2}}\end{aligned}$$

which is also obvious from the figure. We didn't need vectors for this.

$$\begin{aligned}(-\mathbf{w}) \cdot \mathbf{v} &= 2b^2 \\ &= \sqrt{a^2 + b^2} \cdot 2b \cos \phi \\ \cos \phi &= \frac{b}{\sqrt{a^2 + b^2}}\end{aligned}$$

We obtain the same result for $\cos \phi$ as for $\cos \theta$ and then finally

$$\theta = \phi$$

Chapter 18

Vector cross product

Suppose we have two ordinary vectors \mathbf{u} and \mathbf{v} . These must be in \mathbb{R}^3 because the cross-product is only defined for vectors in \mathbb{R}^3 .

Their respective lengths are u and v .

We write the cross-product as

$$\mathbf{u} \times \mathbf{v} = \mathbf{w}$$

The simplest definition is that the magnitude of \mathbf{w} is

$$w = uv \sin \theta$$

The symmetry with the dot product is obvious. Also

$$|\mathbf{u} \times \mathbf{v}|^2 + |\mathbf{u} \cdot \mathbf{v}|^2 = (uv)^2$$

The direction is defined by saying that \mathbf{w} is orthogonal to the plane which contains both \mathbf{u} and \mathbf{v} , and its sign is given by the right-hand rule. Curl the fingers of your right hand around in the direction from \mathbf{u} to \mathbf{v} . Your thumb points in the same direction as \mathbf{w} .

The term $\sin \theta$ means that the cross-product of any vector with itself is zero.

$$\mathbf{a} \times \mathbf{a} = \mathbf{0}$$

To make the notation simpler, we define

$$\mathbf{u} = \langle p, q, r \rangle$$

$$\mathbf{v} = \langle x, y, z \rangle$$

and in order to compute the cross product, we form what looks like a really weird matrix

$$\begin{bmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ p & q & r \\ x & y & z \end{bmatrix}$$

and write its "determinant"

$$\mathbf{u} \times \mathbf{v} = (qz - ry) \hat{\mathbf{i}} + (rx - pz) \hat{\mathbf{j}} + (py - qx) \hat{\mathbf{k}}$$

We can show that the resulting vector is orthogonal to the two starting vectors, \mathbf{u} and \mathbf{v} . Test that by forming the dot product with \mathbf{u} .

$$\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = p(qz - ry) + q(rx - pz) + r(py - qx)$$

The first and fourth terms cancel, the second and fifth terms cancel, and the third and sixth terms also cancel.

So $\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = 0$, and $\mathbf{v} \cdot (\mathbf{u} \times \mathbf{v}) = 0$ as well.

In fact, a very common use for the cross-product is to find the normal vector to a plane in vector calculus.

As an aside, we could have skipped this calculation. The following rule holds for vectors:

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$$

(we will explore triple products below). So

$$\mathbf{u} \cdot (\mathbf{u} \times \mathbf{v}) = (\mathbf{u} \times \mathbf{u}) \cdot \mathbf{v} = 0$$

$$\mathbf{v} \cdot (\mathbf{u} \times \mathbf{v}) = -\mathbf{v} \cdot (\mathbf{v} \times \mathbf{u}) = -(\mathbf{v} \times \mathbf{v}) \cdot \mathbf{u} = 0$$

About the angle

How to show that

$$\mathbf{a} \times \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \sin \theta \hat{\mathbf{n}}$$

where $\hat{\mathbf{n}}$ is perpendicular to \mathbf{a} and \mathbf{b} .

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin \theta$$

According to wikipedia, this is the *definition* of the cross-product, and from this one can derive the expression that we got by setting up our matrix and computing its "determinant." So that is what we are going to do.

I am going to go back to the notation we had before, rather than use subscripts like a_x , etc.

$$\mathbf{u} = \langle p, q, r \rangle$$

$$\mathbf{v} = \langle x, y, z \rangle$$

We proceed from the "determinant" definition of the cross product and show that the length of that vector squared plus the square of the dot product is equal to $u^2 v^2$. By the argument we made above, the magnitude of the cross product is then equal to $uv \sin \theta$.

$$\mathbf{u} \times \mathbf{v} = (qz - ry)\hat{\mathbf{i}} + (rx - pz)\hat{\mathbf{j}} + (py - qx)\hat{\mathbf{k}}$$

$$\begin{aligned} |\mathbf{u} \times \mathbf{v}|^2 &= (qz - ry)^2 + (rx - pz)^2 + (py - qx)^2 \\ &= (qz)^2 - 2qryz + (ry)^2 + (rx)^2 - 2prxz + (pz)^2 + (py)^2 - 2pqxy + (qx)^2 \end{aligned}$$

$$\mathbf{u} \cdot \mathbf{v} = px + qy + rz$$

$$(\mathbf{u} \cdot \mathbf{v})^2 = (px)^2 + (qy)^2 + (rz)^2 + 2pqxy + 2prxz + 2qryz$$

When we add these together, all the terms with cofactor 2 cancel so that leaves

$$\begin{aligned} &|\mathbf{u} \times \mathbf{v}|^2 + (\mathbf{u} \cdot \mathbf{v})^2 \\ &= (qz)^2 + (ry)^2 + (rx)^2 + (pz)^2 + (py)^2 + (qx)^2 + (px)^2 + (qy)^2 + (rz)^2 \end{aligned}$$

rearranging terms

$$= (px)^2 + (py)^2 + (pz)^2 + (qx)^2 + (qy)^2 + (qz)^2 + (rx)^2 + (ry)^2 + (rz)^2$$

$$= (p^2 + q^2 + r^2)(x^2 + y^2 + z^2)$$

$$= |\mathbf{u}|^2 |\mathbf{v}|^2$$

That was tedious, but it we made it.

All of these properties of the cross-product are connected.

$$\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{a} \times \mathbf{b}) = 0$$

$$\mathbf{a} \times \mathbf{b} = \langle qu - rt, rs - pu, pt - qs \rangle$$

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta$$

$$|\mathbf{a} \times \mathbf{b}|^2 + (\mathbf{a} \cdot \mathbf{b})^2 = |\mathbf{a}|^2 |\mathbf{b}|^2$$

Triple products

Suppose we have

$$\mathbf{a} = \langle p, q, r \rangle$$

$$\mathbf{b} = \langle s, t, u \rangle$$

$$\mathbf{c} = \langle x, y, z \rangle$$

And

$$\mathbf{a} \times \mathbf{b} = \langle qu - rt, rs - pu, pt - qs \rangle$$

$$\mathbf{b} \times \mathbf{c} = \langle tz - uy, ux - sz, sy - tx \rangle$$

$$\mathbf{a} \times \mathbf{c} = \langle qz - ry, rx - pz, py - qx \rangle$$

Algebraically

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = p(tz - uy) + q(ux - sz) + r(sy - tx)$$

$$\mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = s(ry - qz) + t(pz - rx) + u(qx - py)$$

$$\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = x(qu - rt) + y(rs - pu) + z(pt - qs)$$

So

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$$

The way to remember this is that these are all the same cyclic permutation.

A much simpler proof is to remember that the cross-product $\mathbf{a} \times \mathbf{b}$ is the area of the parallelogram formed by \mathbf{a} and \mathbf{b} and the *scalar* triple product is the signed volume of the parallelepiped formed by the three vectors. Signed meaning that $\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = -\mathbf{c} \cdot (\mathbf{b} \times \mathbf{a})$ so the area may come out negative, if we order \mathbf{a} and \mathbf{b} differently.

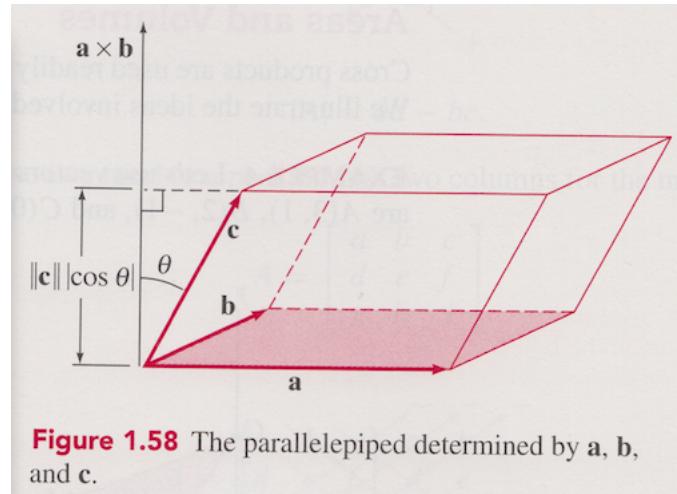


Figure 1.58 The parallelepiped determined by \mathbf{a} , \mathbf{b} , and \mathbf{c} .

Recall that the direction of $\mathbf{a} \times \mathbf{b}$ is perpendicular to both vectors. If we are careful to write the cross-product in the correct order using the right-hand rule, the result of the dot product will always be positive, with the projection of \mathbf{c} onto the cross-product equal to the height of the solid. In particular, for this arrangement, we must write $\mathbf{a} \times \mathbf{b}$, $\mathbf{b} \times \mathbf{c}$, or $\mathbf{c} \times \mathbf{a}$.

It doesn't matter which two vectors we choose as the base of our solid, the volume must come out the same.

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$$

Chapter 19

Point and plane

Construct a plane containing 3 points

Consider three points

$P = (1, 0, 0)$, $Q = (0, 1, 0)$, and $R = (0, 0, 1)$.

Find two vectors in the plane by subtracting the second and third from the first.

$$\begin{aligned}\mathbf{u} &= (1, 0, 0) - (0, 1, 0) \\ &= \langle 1, -1, 0 \rangle \\ \mathbf{v} &= (1, 0, 0) - (0, 0, 1) \\ &= \langle 1, 0, -1 \rangle\end{aligned}$$

Obtain the normal vector by computing the cross product

$$\mathbf{N} = \mathbf{u} \times \mathbf{v} \Rightarrow \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{vmatrix} = 1\hat{\mathbf{i}} + 1\hat{\mathbf{j}} + 1\hat{\mathbf{k}} = \langle 1, 1, 1 \rangle$$

One equation of the plane is then

$$\mathbf{N} \cdot \mathbf{w} = 0$$

for any vector \mathbf{w} in the plane.

Consider a fixed point in the plane (x_0, y_0, z_0) . Then any other point in the plane (x, y, z) yields a vector from the fixed point which, dotted with \mathbf{N} , yields 0

$$\langle x - x_0, y - y_0, z - z_0 \rangle \cdot \langle 1, 1, 1 \rangle = 0$$

$$x - x_0 + y - y_0 + z - z_0 = 0$$

$$x + y + z = x_0 + y_0 + z_0 = d$$

Plugging in any one of the points yields

$$x + y + z = d = 1$$

find the closest point in the plane

Consider any point in space, e.g. $P = (3, 4, 6)$.

Find the point Q on the plane which is closest to P , the point we arrive at by subtracting some fraction of \mathbf{N} from P .

We have a point and a vector

$$Q = P - t\mathbf{N}$$

$$Q = (3, 4, 6) - t\langle 1, 1, 1 \rangle$$

Since Q is in the plane, its components x, y, z satisfy $x + y + z = 1$!

So

$$(3 - t) + (4 - t) + (6 - t) = 1$$

$$13 - 3t = 1$$

$$t = 4$$

$$Q = (-1, 0, 2)$$

Check that Q is in the plane

$$-1 + 0 + 2 = 1$$

and $P - Q$ is parallel to \mathbf{N}

$$P - Q = \langle 4, 4, 4 \rangle$$

that is definitely a multiple of \mathbf{N} .

point where a vector crosses the plane

Where does the vector \mathbf{w} that goes from the origin to point $P = (3, 4, 6)$ hit the plane? Call that point R . Again we have a point and a vector

$$R = (0, 0, 0) + t\mathbf{w} = (0, 0, 0) + t \langle 3, 4, 6 \rangle$$

And again, since R is in the plane, its components x, y, z satisfy $x + y + z = 1$. So

$$3t + 4t + 6t = 1$$

$$t = \frac{1}{13}$$

$$R = \left(\frac{3}{13}, \frac{4}{13}, \frac{6}{13} \right)$$

Notice that the vector $Q - R$ is in the plane, as it should be

$$\begin{aligned} (Q - R) \cdot \mathbf{N} &= ((-1, 0, 2) - \left(\frac{3}{13}, \frac{4}{13}, \frac{6}{13} \right)) \cdot \langle 1, 1, 1 \rangle \\ &= \left\langle \frac{-16}{13}, \frac{-4}{13}, \frac{20}{13} \right\rangle \cdot \langle 1, 1, 1 \rangle = 0 \end{aligned}$$

And, adding the horizontal and vertical components together

$$\begin{aligned} Q - R + P - Q &= P - R = (3, 4, 6) - \left(\frac{3}{13}, \frac{4}{13}, \frac{6}{13} \right) \\ &= \left(\frac{36}{13}, \frac{48}{13}, \frac{72}{13} \right) \end{aligned}$$

the result is parallel to \mathbf{w} .

Lines in space

One way of specifying a line in 3D-space is as the intersection of two planes. Another way is by giving a vector and a point in space. Let's look at these in turn. Suppose we have the following two planes:

$$x + y - z = 7$$

$$2x - 3y + z = 3$$

Since the x, y, z terms are not related by a multiplicative constant, the planes are not parallel, so they will meet in a line, and the solutions consist of all the points on the line. Let's find one solution, at $x = 0$. Then

$$y - z = 7$$

$$-3y + z = 3$$

Adding

$$-2y = 10$$

$$y = -5$$

$$z = y - 7 = -12$$

Our solution $P_0 = (0, -5, -12)$. Now find a second solution, at $z = -3$

$$x + y = 4$$

$$2x - 3y = 6$$

Solving

$$x = 4 - y$$

$$2(4 - y) - 3y = 6$$

$$8 - 5y = 6$$

$$y = \frac{2}{5}$$

$$x = \frac{18}{5}$$

The second point is $P_1 = (18/5, 2/5, -3)$. Now we have two points on the line. Its equation is

$$L = P_0 + t(P_1 - P_0)$$

$$L = (0, -5, -12) + t\left(\frac{18}{5}, \frac{27}{5}, 9\right)$$

We can re-scale the vector that multiplies t to have integer components (or length 1, or whatever we wish). Why not multiply by 5/9?

$$L = (0, -5, -12) + t(2, 3, 5)$$

There is another way to do this problem that might be a little easier. Consider that the equation of the first plane gives its normal vector n_1 as

$$n_1 = \langle 1, 1, -1 \rangle$$

Similarly the normal vector to the second plane is n_2

$$n_2 = \langle 2, -3, 1 \rangle$$

Now, the vector that is parallel to the line of intersection is orthogonal to both n_1 and n_2 (Do you see why?) So we compute the cross-product:

$$\begin{aligned} n_1 \times n_2 &= \begin{vmatrix} i & j & k \\ 1 & 1 & -1 \\ 2 & -3 & 1 \end{vmatrix} \\ &= -2i - 3j - 5k \end{aligned}$$

Multiplying by -1 gives what we obtained above.

Chapter 20

Headlight problem

This chapter assumes a bit of knowledge about vectors. It's basically geometry so I decided to put it in an appendix. If you don't know about vectors you could learn more by going to the calculus book (link in the Introduction). Very little is required except for the idea of the dot product.

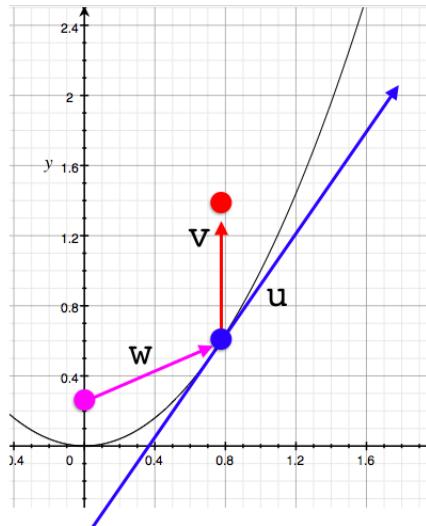
The reflective property of the parabola asserts that if a light ray emitted from the focus bounces off any point of the parabola, it then travels off in the vertical direction.



Snell's law for reflection says that the angle of incidence and reflection to the inside surface of the parabola must be equal. It is curious that this law has Snell's name on it, since the fact was known to Euclid, and Heron had a proof of it. The proof

depends on the assumption that light travels the shortest path between two points.

In any case, applying that law to this problem, the angle of incidence is the angle of the magenta vector \mathbf{w} with the tangent vector \mathbf{u} . This is equal to the angle of reflection, the angle of the tangent \mathbf{u} with the vertical vector \mathbf{v} .



We assert that there exists a point on the y -axis (the focus, colored magenta), with the property that when we draw a vector to any point on the parabola, the angle that this vector makes with the tangent to the parabola is equal to the angle the tangent makes with the vertical.

Let the distance of this point from the origin be p . Then

$$\mathbf{w} = \langle x, ax^2 - p \rangle$$

The tangent has slope $2ax$ so

$$\mathbf{u} = \langle 1, 2ax \rangle$$

Scale the vertical to be a unit vector

$$\mathbf{v} = \langle 0, 1 \rangle$$

By the definition of the dot product, the cosine of the angle between \mathbf{w} and \mathbf{u} is

$$\frac{\mathbf{w} \cdot \mathbf{u}}{u w}$$

By the equal angle constraint, this is equal to the cosine of the angle between \mathbf{u} and \mathbf{v}

$$\frac{\mathbf{u} \cdot \mathbf{v}}{u v} = \frac{\mathbf{w} \cdot \mathbf{u}}{u w}$$

Since $v = 1$ we have

$$w (\mathbf{u} \cdot \mathbf{v}) = \mathbf{w} \cdot \mathbf{u}$$

That's the important logic of the solution.

Now it's just algebra: The length of \mathbf{w} is

$$w = \sqrt{x^2 + (ax^2 - p)^2}$$

while

$$\mathbf{u} \cdot \mathbf{v} = 2ax$$

$$\mathbf{w} \cdot \mathbf{u} = x + 2ax(ax^2 - p)$$

So

$$\begin{aligned} w (\mathbf{u} \cdot \mathbf{v}) &= \mathbf{w} \cdot \mathbf{u} \\ \sqrt{x^2 + (ax^2 - p)^2} (2ax) &= x + 2ax(ax^2 - p) \end{aligned}$$

Divide by $2ax$:

$$\sqrt{x^2 + (ax^2 - p)^2} = \frac{1}{2a} + (ax^2 - p)$$

Square both sides

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

A nice cancellation:

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

We can also cancel the x^2 :

$$0 = \frac{1}{(2a)^2} + \frac{1}{a}(-p)$$

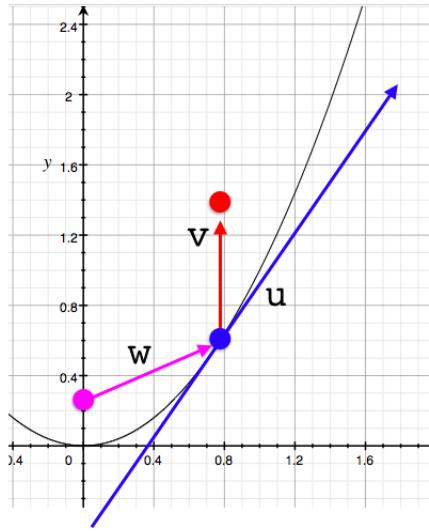
and finally cancel an a :

$$0 = \frac{1}{4a} - p$$

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

The point $(0, 1/4a)$ is, as we saw before, the focus of the parabola.

Since p is independent of x , this property holds for every point on the parabola.



An alternative, more geometric approach is to note that the angle the vector **u** makes with the vertical at (x, ax^2) is equal to the angle **u** makes with the y -axis (just off the image to the bottom).

This angle is equal to the angle between **w** and **u** if and only if the triangle is isosceles, that is, if length of the vector **w** is equal to the distance between $(0, p)$ and the intersection of **u** with the y -axis.

We start by exploring the properties of a line through the point (x, ax^2) with slope equal to $2ax$.

From this point on, the point on the parabola is *fixed*. We want to write an equation for a line with the same slope as the parabola at this point, the same slope as the vector **u**.

We will be re-using x as a variable. To reduce confusion, label the fixed value at the point as \hat{x} , so then $\hat{y} = a\hat{x}^2$, and the slope is $2a\hat{x}$.

The point-slope formula for the line is

$$2a\hat{x} = \frac{\Delta y}{\Delta x} = \frac{y - \hat{y}}{x - \hat{x}} = \frac{y - a\hat{x}^2}{x - \hat{x}}$$

The intersection with the y -axis occurs at $y = 0$ so there

$$2a\hat{x} = \frac{-a\hat{x}^2}{x - \hat{x}}$$

$$2 = \frac{-\hat{x}}{x - \hat{x}}$$

$$2x - 2\hat{x} = -\hat{x}$$

$$x = \frac{\hat{x}}{2}$$

The intersection of \mathbf{u} with the x -axis is at $\hat{x}/2$.

For the intersection with the y -axis, $x = 0$ and then

$$2a\hat{x} = \frac{y - a\hat{x}^2}{-\hat{x}}$$

$$-2a\hat{x}^2 = y - a\hat{x}^2$$

$$y = -a\hat{x}^2$$

What we've discovered is that the point of intersection is the same distance below the x -axis as our point on the parabola $(\hat{x}, a\hat{x}^2)$ is above it. We could have used congruent triangles proceeding from the discovery above that the intersection of \mathbf{u} with the x -axis is at $\hat{x}/2$.

Our goal is to show that the triangle is isosceles:

$$a\hat{x}^2 + p = w$$

$$a\hat{x}^2 + p = \sqrt{\hat{x}^2 + (a\hat{x}^2 - p)^2}$$

$$(a\hat{x}^2 + p)^2 = \hat{x}^2 + (a\hat{x}^2 - p)^2$$

Continuing

$$a^2\hat{x}^4 + 2ap\hat{x}^2 + p^2 = \hat{x}^2 + a^2\hat{x}^4 - 2ap\hat{x}^2 + p^2$$

Does this look familiar?

Cancel two terms

$$2ap\hat{x}^2 = \hat{x}^2 - 2ap\hat{x}^2$$

$$4ap\hat{x}^2 = \hat{x}^2$$

$$4ap = 1$$

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

And we already proved this is true, if the magenta point we start from is the focus.

Hence the lengths are equal, the triangle is isosceles, and the corresponding angles are equal. The point we've been using is just the focus.

Chapter 21

Rotation

Rotation can make life difficult when working with conic sections.

We typically draw curves symmetric about an axis, usually the y -axis, but sometimes x - for variety. However, there are other possibilities.

Consider the parabola. Usually, one opens up or down. Sometimes it is rotated to open right or left. These are obtained by having an equation like

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

with $x = f(y)$.

To work a problem like this, I find the easiest thing to do is to switch x for y , solve the problem, and switch back at the end.

But it is also possible to rotate through a different angle, like 45° . What happens then? Well, basically we replace x and y with rotated coordinates u and v (or x' and y' , whatever):

For example we can use:

$$x = u \cos \theta - v \sin \theta$$

$$y = u \sin \theta + v \cos \theta$$

I derived these [here](#).

One thing that may puzzle you is a particular sign difference:

$$r_{cw} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \quad r_{ccw} = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}$$

What's going on is that rotation may be clockwise or counter-clockwise.

(Another thing that can happen is that a cw rotation of the coordinate system is the same as the ccw rotation of the points. So it depends on the derivation.)

You can test by rotating the unit vector $(1, 0)$. Let $t = 90$ degrees counter-clockwise and then

$$r_{ccw} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$r_{ccw}(1, 0) = (0, 1)$, turning $\hat{\mathbf{i}}$ into $\hat{\mathbf{j}}$, which checks.

It is just a matter of notation whether we write the matrix form or something like

$$\begin{aligned} x &= u \cos \theta - v \sin \theta \\ y &= u \sin \theta + v \cos \theta \end{aligned}$$

With 45° , $\sin \theta = \cos \theta = 1/\sqrt{2}$. Let

$$k = \sin \theta = \cos \theta = 1/\sqrt{2}$$

Consider the parabola

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

and substitute for x and y as given above

$$\begin{aligned} ku + kv &= a(ku - kv)^2 + b(ku - kv) + c \\ u + v &= ak(u^2 - 2uv + v^2) + b(u - v) + \frac{c}{k} \end{aligned}$$

Now, we might attempt to solve this for v in terms of u , but there is a new term $-2uv$ which mixes up u and v .

Upon encountering a problem with mixed xy and other powers, the best thing is to rotate to get rid of the xy stuff, and we'll see how to do that in the next section.

For a second example, consider the hyperbola:

$$x^2 - y^2 = 2$$

This curve has $x = \sqrt{2}$ when $y = 0$, (and no y value satisfies the equation when $x = 0$). It opens to left and right. $\sqrt{2}$ is the distance from the origin to the vertex.

Now we want to turn the curve by 45° ccw.

$\sin \pi/4 = \cos \pi/4 = 1/\sqrt{2}$. So the matrix is

$$r_{ccw} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

which amounts to writing

$$x = \frac{1}{\sqrt{2}}(u - v), \quad y = \frac{1}{\sqrt{2}}(u + v)$$

so

$$\begin{aligned} x^2 - y^2 &= 2 \\ (u - v)^2 - (u + v)^2 &= 4 \\ u^2 + 2uv + v^2 - u^2 + 2uv - v^2 &= 4 \\ uv &= 1 \end{aligned}$$

This certainly looks familiar!

A plot will show that $(1, 1)$ is on the curve at the point of closest approach. The distance from the origin to the vertex is thus $\sqrt{2}$, which is the same as the one we started with above, $x^2 - y^2 = 2$.

general problem

The most general equation for a parabola, ellipse or hyperbola is

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

This applies to rotated versions of all three.

Kline says (Chapter 7) to consider a rotation through an angle θ . I will use t for θ . We wrote above

$$\begin{aligned} x &= u \cos t - v \sin t \\ y &= u \sin t + v \cos t \end{aligned}$$

First compute the products:

- o $x^2 = u^2 \cos^2 t - 2uv \sin t \cos t + v^2 \sin^2 t$
- o $xy = u^2 \sin t \cos t + uv \cos^2 t - uv \sin^2 t - v^2 \sin t \cos t$
- o $y^2 = u^2 \sin^2 t + 2uv \sin t \cos t + v^2 \cos^2 t$

Now try substituting into the general equation (I know, it's a mess). We collect the coefficients for all the terms u^2 , uv , v^2 , etc., separately:

- o $[A \cos^2 t + B \sin t \cos t + C \sin^2 t] u^2$
- o $[-2A \sin t \cos t + B \cos^2 t - B \sin^2 t + 2C \sin t \cos t] uv$
- o $[A \sin^2 t - B \sin t \cos t + C \cos^2 t] v^2$
- o $[D \cos t + E \sin t] u$
- o $[-D \sin t + E \cos t] v$

We don't need most of this.

The clever insight is that we can choose the angle t so as to eliminate the coefficient of the term that mixes u and v : namely uv .

$$-2A \sin t \cos t + B \cos^2 t - B \sin^2 t + 2C \sin t \cos t = 0$$

Remember those sum of angles formulas!

$$\cos^2 t - \sin^2 t = \cos 2t$$

$$2 \sin t \cos t = \sin 2t$$

So

$$-A \sin 2t + B \cos 2t + C \sin 2t = 0$$

$$(C - A) \sin 2t + B \cos 2t = 0$$

giving

$$\cot 2t = \frac{A - C}{B}$$

$$\tan 2t = \frac{B}{A - C}$$

example

Consider again

$$xy = 1$$

Here A and C are zero, while $B = 1$. What angle's tangent is not defined? $\pi/2$. As $2t$ approaches $\pi/2$, its tangent approaches ∞ . So the value of t we seek is $t = \pi/4$.

We go back and compute the coefficients for all the other terms. Since only $B \neq 0$ and since $\sin \pi/4 = \cos \pi/4 = 1/\sqrt{2}$, we get

$$\begin{aligned} & [\frac{A}{2} + \frac{B}{2} + \frac{C}{2}] u^2 + [\frac{A}{2} - \frac{B}{2} + \frac{C}{2}] v^2 = 1 \\ & = \frac{u^2}{2} - \frac{v^2}{2} = 1 \\ & u^2 - v^2 = 2 \end{aligned}$$

which is the equation of a rectangular hyperbola opening left and right.

test

Suppose you run into a general conic equation with some version of

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

Ask these questions to decide what you have:

- **Are both variables squared?**

No: It's a parabola.
Yes: Go to the next test....

- **Do the squared terms have opposite signs?**

Yes: It's an hyperbola.
No: Go to the next test....

- **Are the squared terms multiplied by the same number?**

Yes: It's a circle.
No: It's an ellipse.

Kline goes through the effort of showing that, after rotating to a standard orientation, *every* equation of the general form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

can be translated to the origin to give a standard parabola, circle, ellipse or hyperbola.

Not every quadratic equation gives a conic. Some are "degenerate".

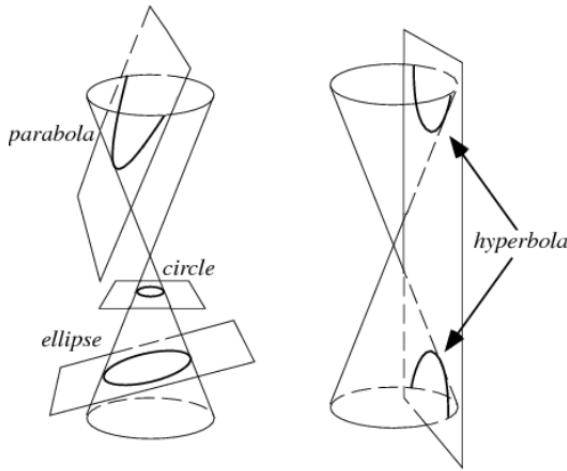
For example, having done all the right manipulations, we might end up with something like

$$A'(x - h)^2 + B'(y - k)^2 = 0$$

which has only $x = h$ and $y = k$ as a solution. It's a point. And if A , C and F are all negative: there is no solution in the real numbers.

conic sections

Everyone learns in high school that the conic sections can be obtained by slicing a double cone with a plane and taking the points that belong to both. This can get complicated for several reasons, which is why they don't usually do examples, even in basic calculus.



The inclination of the plane must match that of the cone in order to obtain a parabola. Take the dot product of the normal vector to the plane with the unit vector for the z -axis, divide by the length of \mathbf{N} , and that's $\cos \phi$, the angle made with the vertical. If we call θ the angle with the xy -plane, that's the complement of ϕ . Which has to match the inclination of the cone, namely $H/R = \cos \theta$.

The distance between the point of closest approach to the plane and the vertex of the cone will change the shape factor. As you go up the cone, the curvature of the level curves of the cone becomes shallower.

And the orientation of the plane, the direction in which its normal vector points, will determine whether the final result has any terms that mix x and y . For our example, we pick a normal vector that aligns with the y -axis. We want $\mathbf{N} \cdot \hat{\mathbf{i}} = 0$, which will happen if the x -component of \mathbf{N} is zero.

Here is a simple example.

The level curves of a cone are

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

and the equation of the cone is $z = kr$ where $k = H/R$ is a constant. For simplicity, suppose $k = 1$.

Now, let's have a plane like

$$z = y + 1$$

This plane has normal vector

$$\mathbf{N} = \langle 0, -1, 1 \rangle$$

so the x -axis lies in the plane because $\mathbf{N} \cdot \hat{\mathbf{i}} = 0$. Another vector orthogonal to both and also in the plane is $\langle 0, 1, 1 \rangle$.

The normal vector to the cone depends on where you are, but if you are at $x = 0, y = 1, z = 1$ then it would be

$$\mathbf{N} = \langle 0, 1, -1 \rangle$$

In the yz -plane it points down at a 45 degree angle. The two normal vectors are orthogonal, so if there is a solution it should be a parabola.

We can see that there should be a solution, because the plane intersects the y -axis at $y = -1$ ($z = 0$) and the z -axis at $z = 1$. If you draw a sketch, one point is outside the cone and the other inside it, so the plane must cut the cone.

Every point on the intersection of the plane and the cone satisfies both equations:

$$\begin{aligned}\sqrt{x^2 + y^2} &= 1 + y \\ x^2 + y^2 &= 1 + 2y + y^2 \\ x^2/2 &= y + 1/2\end{aligned}$$

This is a parabola but it is *not* the parabola formed by the intersection. It is the projection of that intersection onto the xy -plane. You can tell because it has no z in it.

Such projections are linear transformations, which simply amount to rescaling of the variables x to x' and y to y' (in this case only the latter) without changing the nature of the curve—a parabola is still a parabola.

However, an ellipse may become a circle, and vice-versa.

In this case, the normal vector forms an angle of 45 degrees with the vertical z -axis since

$$\cos \theta = \frac{\langle 0, -1, 1 \rangle \cdot \langle 0, 0, 1 \rangle}{\sqrt{1+1} \sqrt{1}} = \frac{1}{\sqrt{2}}$$

This is the factor by which the actual curve is stretched compared to the projection in the plane.

For the general problem to find the equation of the parabola drawn in our inclined plane, we would need to rotate all the points on the curve using angles obtained from the normal vector. We want to tilt \mathbf{N} so that it points straight up and has its magnitude unchanged.

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

In 3D we could rotate points (or the coordinate system) first with respect to the xy -plane (ignoring z) using the standard transformation with this matrix

$$\begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

This is the same rotation that we had before — it leaves the z -coordinate unchanged. The relevant t is calculated from the x and y components of \mathbf{N} using $t = \tan^{-1} y/x$. Then use the given matrix, or perhaps switch the signs on the sine.

After \mathbf{N} has been rotated so that it lies along either the x - or y -axis, then rotate in the xz -plane or yz -plane until \mathbf{N} is vertical.

Having done this, I believe there should be no mixed terms containing xy , so we won't need to rotate to remove those, as done before.

Part VI

Theta and r

Chapter 22

Polar coordinates

In polar coordinates points are plotted in terms of distance from the origin, r and the angle θ that this ray makes with the positive x -axis. Converting from x, y to r, θ is pretty easy:

$$x = r \cos \theta$$

$$y = r \sin \theta$$

To go the other way, use Pythagoras to write

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

$$\theta = \tan^{-1}\left(\frac{y}{x}\right), \quad x \neq 0$$

In polar coordinates, as in Cartesian (xy) coordinates, the equation of a circle depends on whether it is at the origin or not. If it is at the origin, then something like

$$r = 3$$

defines the graph. But if it's away from the origin, then the equations are of the form:

$$r = a \cos \theta + b \sin \theta$$

Let's do a derivation and then look at examples. We can manipulate these equations to go back to Cartesian coordinates.

$$r = 2h \cos \theta + 2k \sin \theta$$

The reason for choosing these particular coefficients will become clear shortly. Substitute x and y .

$$r = 2h \cdot \frac{x}{r} + 2k \cdot \frac{y}{r}$$

so

$$\begin{aligned} r^2 &= 2hx + 2ky \\ x^2 + y^2 &= 2hx + 2ky \\ [x^2 - 2hx] + [y^2 - 2ky] &= 0 \end{aligned}$$

complete both squares

$$\begin{aligned} [x^2 - 2hx + h^2] + [y^2 - 2ky + k^2] &= h^2 + k^2 \\ (x - h)^2 + (y - k)^2 &= h^2 + k^2 \end{aligned}$$

That is, for an equation of the form

$$r = 2h \cos \theta + 2k \sin \theta$$

the origin is at (h, k) and the radius is

$$r = \sqrt{h^2 + k^2}$$

If the equation contains only $\sin \theta$ then compute k equal to one-half the coefficient of $\sin \theta$, with the origin at $(0, k)$ and radius $r = k$.

Similarly, if the equation contains only $\cos \theta$ then compute h equal to one-half the coefficient of $\cos \theta$, with the origin at $(h, 0)$ and radius $r = h$.

examples

For example

$$r = 3 \sin \theta$$

is a circle centered at $(0, 3/2)$, with a radius of $3/2$ (it passes through the origin and the point $(x = 0, y = 3)$). All such circles (with just one of $\sin \theta$ or $\cos \theta$) have this property.

And

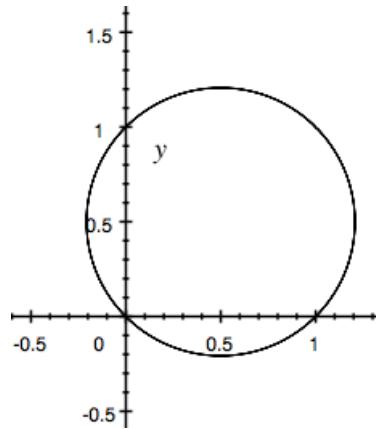
$$r = \sin \theta + \cos \theta$$

is a circle centered at $(1/2, 1/2)$ with a radius squared:

$$r^2 = h^2 + k^2 = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$r = \frac{1}{\sqrt{2}}$$

We see that all circles of this form pass through the origin.



other conic sections

Parabolas look like this:

$$r = \frac{1}{1 \pm a \sin \theta}$$

$$r = \frac{1}{1 \pm a \cos \theta}$$

The ones with $\sin \theta$ open up and down, the others open left and right. If the sign of the a term is negative, the parabola opens up or to the right.

The general formulas are

$$r = \frac{ep}{1 \pm e \sin \theta} = p \frac{1}{1/e + \sin \theta}$$

$$r = \frac{ep}{1 \pm e \cos \theta} = p \frac{1}{1/e + \cos \theta}$$

where e is the eccentricity ($e = 1$ for a parabola). If $e < 1$ then we have an **ellipse**.

parabola

Consider

$$r = \frac{2}{1 + \sin \theta}$$

Plot it to see. Or convert to Cartesian coordinates:

$$r + r \sin \theta = 2$$

$$\frac{y}{r} = \sin \theta$$

$$r + y = 2$$

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

$$4 - 4y + y^2 = x^2 + y^2$$

$$y - 1 = -\frac{1}{4}x^2$$

ellipse

If we measure r and θ from a focus, then

$$x = c + r \cos \theta$$

$$y = r \sin \theta$$

one can derive a formula for the ellipse in terms of r, θ

$$r = \frac{ep}{1 \pm e \cos \theta}$$

(We talk more about this [here](#)). For example if

$$r = \frac{1}{1 + e \cos \theta}$$

with $0 < e < 1$, we will get an ellipse.

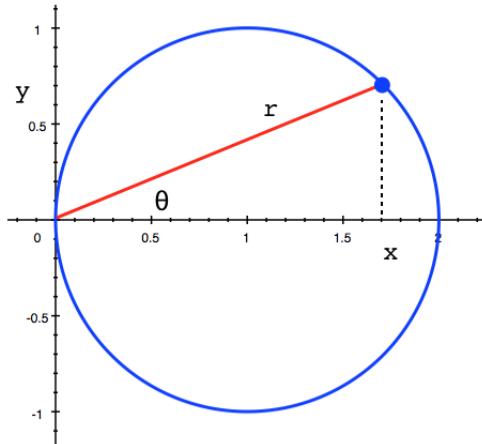
Chapter 23

Polar conics

circle

A very simple circle in polar coordinates is $r = a$. There is no θ -dependence when the circle has its center at the origin.

For a circle of radius a centered at $(a, 0)$ then



$$\begin{aligned} a^2 &= (x - a)^2 + y^2 \\ x^2 - 2ax + y^2 &= 0 \end{aligned}$$

Always, $x = r \cos \theta$ and $y = r \sin \theta$ so

$$r^2(\sin^2 \theta + \cos^2 \theta) - 2ar \cos \theta = 0$$

$$r^2 - 2ar \cos \theta = 0$$

$$r = 2a \cos \theta$$

If the center of the circle is on the y -axis the equation is similar but with $\sin \theta$. A more general equation is

$$r = 2h \cos \theta + 2k \sin \theta$$

which is a circle that touches the origin, and has its center at (h, k) .

The most general equation is with the circle anywhere in the plane. If we remember to specify the center at (s, ϕ) in *radial* coordinates, then the law of cosines easily yields

$$r^2 + s^2 - 2rs \cos(\theta - \phi) = a^2$$

reverse

Start from

$$r = 2h \cos \theta + 2k \sin \theta$$

Always, $x = r \cos \theta$ and $y = r \sin \theta$ so

$$r = 2h \frac{x}{r} + 2k \frac{y}{r}$$

$$r^2 = 2hx + 2ky$$

$$x^2 + y^2 = 2hx + 2ky$$

Easily rearrange and complete the square:

$$x^2 - 2hx + h^2 + y^2 - 2ky + k^2 = h^2 + k^2$$

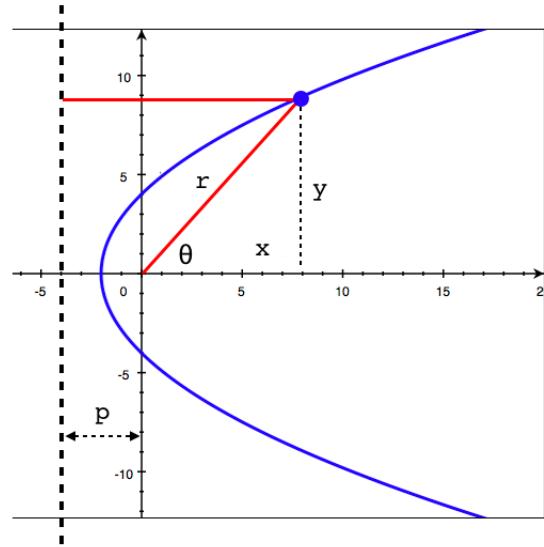
$$(x - h)^2 + (y - k)^2 = h^2 + k^2$$

For a circle touching the origin, $h^2 + k^2 = a^2$

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

parabola

To derive the equation for a parabola in polar coordinates it is convenient to rotate from the standard orientation by 90 degrees CW. In this way θ will have its usual relationship with the x -axis.



The origin of coordinates is placed at the focus, the distance to the vertex for this parabola is 2, and the distance from the origin to the directrix is $p = 4$.

Note that in Cartesian coordinates, this parabola will be of the form $x = ay^2$ because of the rotation.

The distance from the focus to a general point (x, y) is just r . The distance from the directrix to the point is $p + x$. The geometric constraint gives simply:

$$r = p + x$$

We make the standard substitution $x = r \cos \theta$.

$$r = p + r \cos \theta$$

Some rearrangement gives the standard equation

$$r = \frac{p}{1 - \cos \theta}$$

For a vertically oriented parabola we would have $\sin \theta$ instead.

reverse

To go back to Cartesian coordinates, reverse the substitution for x :

$$\begin{aligned} r &= \frac{p}{1 - x/r} \\ r - x &= p \\ r^2 &= (x + p)^2 \end{aligned}$$

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

$$\begin{aligned} x^2 + y^2 &= x^2 + 2px + p^2 \\ y^2 &= 2px + p^2 \\ \frac{1}{2p}y^2 &= x + \frac{p}{2} \end{aligned}$$

This looks unusual. However, the equation that was actually plotted was $r = 4/(1 + \cos \theta)$ ($p = 4$).

Note: here we have used p as the distance from the focus to the directrix, which is twice the distance to the vertex. If we call the latter distance c , the $p = 2c$. Previously we showed that $4ac = 1$, so $a = 1/4c$. Thus we obtain $a = 1/8$:

$$\frac{1}{8}y^2 = x + 2$$

This shape factor matches the plot (four units above the axis (at $x = 0$ is two units to the right of the vertex) and the vertex is at $(-2, 0)$. a is unusually small, the reason is so the parabola will open quickly, giving room to put all the labels in the diagram.

ellipse

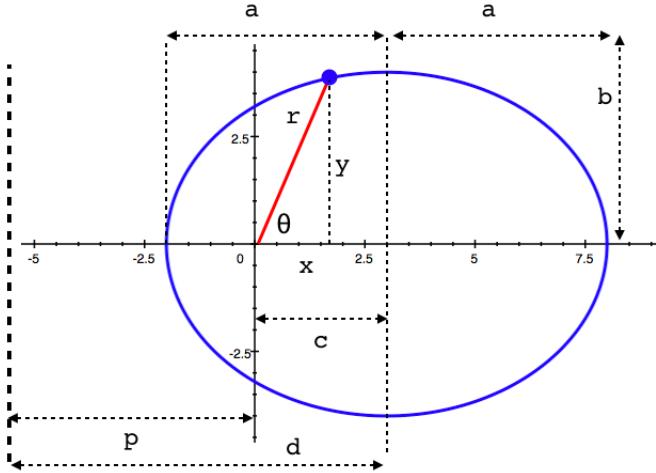
From the geometry of the ellipse, with the center at the origin, it is fairly easy to show that

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

and derive the equation of the ellipse in Cartesian coordinates:

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

To derive the equation for an ellipse in polar coordinates it is convenient to shift the origin of coordinates to be the left focus of the ellipse at $(-c, 0)$.



The ellipse plotted here has $a = 5, b = 4$ and so $c = \sqrt{a^2 - b^2} = 3$. It has been shifted so the focus at $(-3, 0)$ is the origin of coordinates.

The eccentricity e is defined by the geometric constraint (next), which can be shown to be equivalent to $c/a = 0.6$.

Let p be the distance from the focus to the directrix, and let d be the distance from the directrix to the center of the ellipse.

The ellipse can be defined by its **geometric constraint**.

This says that for any point on the ellipse, the ratio of the distance from the focus (and here, the origin) to the point (that is, r), when divided by the distance from the point to the directrix, $x+p$, is equal to a constant, which we will call the eccentricity e .

$$\frac{r}{x+p} = \frac{r}{p+r \cos \theta} = e$$

$$r = e(p + r \cos \theta)$$

We simply rearrange to isolate r

$$r(1 - e \cos \theta) = ep$$

$$r = \frac{ep}{1 - e \cos \theta}$$

reverse

Going back is more complicated for the ellipse. Reverse the substitution $x/r = \cos \theta$.

$$r(1 - ex/r) = ep$$

$$r - ex = ep$$

$$r = ex + ep$$

There's a *magic* substitution that we will justify below:

$$ep = a(1 - e^2)$$

Using that, we have

$$r = ex + a(1 - e^2)$$

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

$$x^2 + y^2 = e^2 x^2 + 2exa(1 - e^2) + a^2(1 - e^2)^2$$

Combine cofactors for x^2 , obtaining $(1 - e^2)$ and then divide through by $(1 - e^2)$:

$$x^2 + \frac{y^2}{1 - e^2} = 2exa + a^2(1 - e^2)$$

Complete the square for x by adding $(ea)^2$ to both sides

$$\begin{aligned} x^2 - 2exa + (ea)^2 + \frac{y^2}{1 - e^2} &= a^2(1 - e^2) + (ea)^2 \\ (x - ea)^2 + \frac{y^2}{1 - e^2} &= a^2(1 - e^2) + (ea)^2 \end{aligned}$$

We asserted that $ea = c$. Simplify the right-hand side at the same time:

$$(x - c)^2 + \frac{y^2}{1 - e^2} = a^2$$

This is great, because we need to shift the origin of coordinates back to the center of the ellipse by exactly this amount.

Unfortunately, I have not discovered any way to make that derivation simpler.

solve for $1 - e^2$

To deal with $1 - e^2$, recall that the basic geometry says

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

$$1 - \left(\frac{c}{a}\right)^2 = \frac{b^2}{a^2}$$

Since $c = ea$

$$1 - e^2 = \frac{b^2}{a^2}$$

so what we had simplifies as the inverse of that times y^2

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

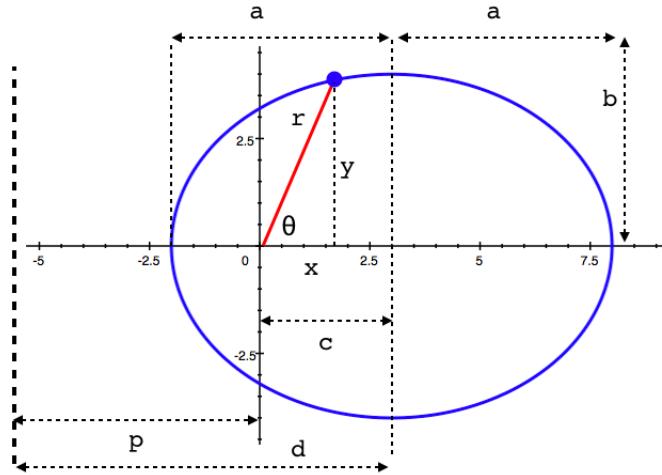
$$\frac{(x - c)^2}{a^2} + \frac{y^2}{b^2} = 1$$

Exactly what we want. Furthermore, we can view this derivation in reverse as a proof of $c = ea$ for an ellipse with this equation in Cartesian coordinates, shifted out to focus $(-c, 0)$.

Now let's explain the substitution:

$$ep = a(1 - e^2)$$

It is easiest to start by finding an expression for d , then getting e and p .



solve for d and e

Applying the geometric constraint to the extreme left end:

$$\frac{a - c}{d - a} = e$$

At the very top of the ellipse the distance to the focus is $\sqrt{b^2 + c^2}$ but this is also just a , which means

$$\frac{a}{d} = e = \frac{a - c}{d - a}$$

so

$$ad - a^2 = ad - cd$$

thus

$$d = \frac{a^2}{c}$$

One can also obtain this result by equating the ratios for the left and right ends of the ellipse.

Notice that the ratio a/d obeys the geometric constraint:

$$\frac{a}{d} = e = \frac{ac}{a^2} = \frac{c}{a}$$

We have proved $ae = c$, using only the geometry.

A longer, but pretty, proof is to start from the ratio for the extreme right end:

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

substitute $d = a^2/c$

$$= \frac{a + c}{a^2/c + a}$$

Multiply top and bottom by $1/a$

$$e = \frac{1 + c/a}{a/c + 1}$$

Put top and bottom over common denominators

$$e = \frac{(a + c)/a}{(a + c)/c} = \frac{c}{a}$$

solve for p

$$\begin{aligned} p &= d - c = \frac{a^2}{c} - c \\ &= \frac{a^2 - c^2}{c} = \frac{b^2}{c} \end{aligned}$$

finally

$$\begin{aligned} pe &= \frac{b^2}{c} \cdot \frac{c}{a} = \frac{b^2}{a} \\ &= \frac{a^2 - c^2}{a} \\ &= \frac{a^2 - e^2 a^2}{a} \\ &= a(1 - e^2) \end{aligned}$$

which is the special substitution we used.

summary of the summary

The circle (touching the origin), parabola (rotated to the right), and the ellipse are, in order:

$$\begin{aligned} r &= 2h \cos \theta + 2k \sin \theta \\ r &= \frac{p}{1 - \cos \theta} \\ r &= \frac{ep}{1 - e \cos \theta} \end{aligned}$$

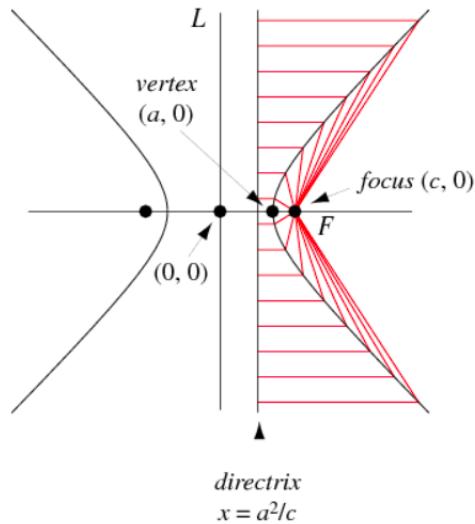
In the last case, note that $0 < e < 1$, so the parabola is the same but with $e = 1$.

Chapter 24

Polar hyperbola

We can also use a focus-directrix approach to the definition of the hyperbola. It will turn out to give the same formula as that of the ellipse and the parabola, but there are a few twists and turns along the way.

<http://mathworld.wolfram.com/Hyperbola.html>



The equation will turn out to be similar to that for the ellipse and parabola

$$r = \frac{ep}{1 - e \cos \theta}$$

except that $e > 1$, as we will see.

First, though, recall the equation that we derived before, namely

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

meaning of a and b

This gives a hyperbola of the type graphed above, opening "east-west". As we used in the derivation, c is the distance from the origin to each focus. Although the value of x is never 0, the value of y can be, and at those two points we have

$$x^2 = a^2$$

$$x = \pm a$$

So a is the horizontal distance to points on the curve along the x -axis.

As for b , we rearrange the basic equation

$$b^2 x^2 - a^2 y^2 = a^2 b^2$$

asymptotes

Associated with the hyperbola is a pair of lines called asymptotes. Their equation is

$$b^2 x^2 - a^2 y^2 = 0$$

Factoring

$$(bx + ay)(bx - ay) =$$

which has solutions when

$$y = \pm \frac{b}{a} x$$

For a hyperbola in standard orientation like this, these lines go through the origin.

eccentricity

Notice that, unlike the ellipse, $a < c$. We define the eccentricity with the same equation as for the ellipse

$$ea = c$$

but now realize that for a hyperbola $e > 1$.

Recall that we defined

$$b^2 = c^2 - a^2, \quad (c > a)$$

Hence

$$b^2 = (ea)^2 - a^2 = a^2(e^2 - 1)$$

whereas before for the ellipse we had

$$b^2 = a^2(1 - e^2)$$

directrix

For the directrix we will assume the answer, and then show that it leads to the desired properties. From the diagram above we read that on the directrix

$$x = \frac{a^2}{c}$$

That is, the distance d from the y -axis is equal to

$$d = \frac{a^2}{c}$$

and since $c = ea$

$$d = \frac{a}{e}$$

which is just what we had before with the ellipse:

$$ea = c$$

$$ed = a$$

(except that now with the hyperbola $e > 1$ and $c > a > d$ whereas before with the ellipse $e < 1$ and $d > a > c$).

On the x -axis the distance from the focus to the curve is $c - a$ and from the curve to the directrix is $a - d$. We consider the ratio of the two distances

$$\begin{aligned} \frac{c - a}{a - d} &= \frac{ea - a}{a - a/e} \\ &= \frac{a(e - 1)}{a(1 - 1/e)} \end{aligned}$$

$$= \frac{(e - 1)}{(1 - 1/e)}$$

Multiply by e on top and bottom:

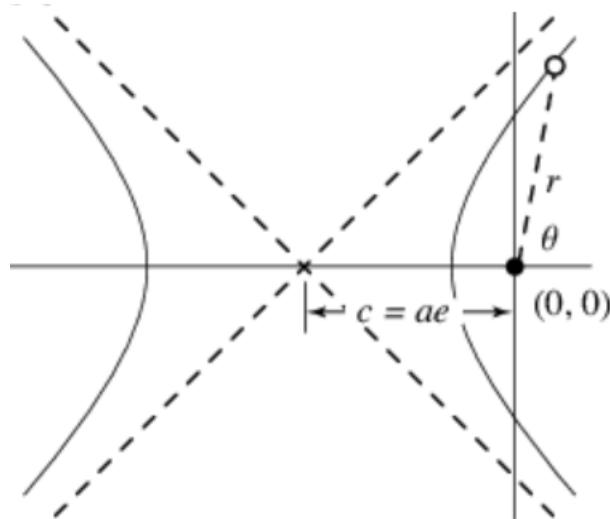
$$= e \frac{(e - 1)}{(e - 1)} \\ = e$$

definition of p

As before, in a similar way to the directrix $d = a^2/c$, we define the focal parameter p as

$$\begin{aligned} p &= \frac{b^2}{c} \\ &= \frac{c^2 - a^2}{c} \\ &= \frac{e^2 a^2 - a^2}{c} \\ &= \frac{a^2(e^2 - 1)}{c} \\ &= \frac{a(e^2 - 1)}{e} \end{aligned}$$

polar coordinates



The geometric constraint is

$$\frac{PF}{PD} = e$$

where PF is just r and the problem is to evaluate the length of PD .

$$PD = r \cos \theta + (c - d)$$

Hence we have that

$$\begin{aligned} e &= \frac{r}{r \cos \theta + (c - d)} \\ er \cos \theta + e(c - d) &= r \\ r(e \cos \theta - 1) &= -e(c - d) \\ r &= \frac{e(c - d)}{1 - e \cos \theta} \end{aligned}$$

The numerator

$$\begin{aligned} e(c - d) &= e\left(ea - \frac{a}{e}\right) \\ &= a(e^2 - 1) \\ &= ep \end{aligned}$$

So finally we obtain:

$$r = \frac{ep}{1 - e \cos \theta}$$

as the equation of a standard hyperbola in polar coordinates.

We recovered the same equation for all three conic sections: parabola, ellipse and hyperbola. The only difference is the value of e . Here $e > 1$, for the parabola $e = 1$ and for the ellipse $e < 1$.

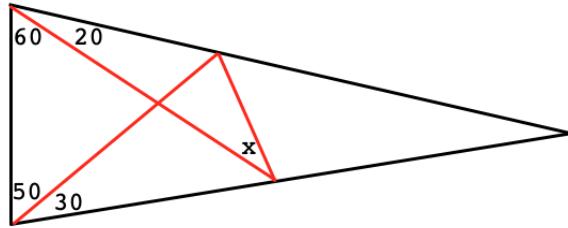
Part VII

Advanced problems

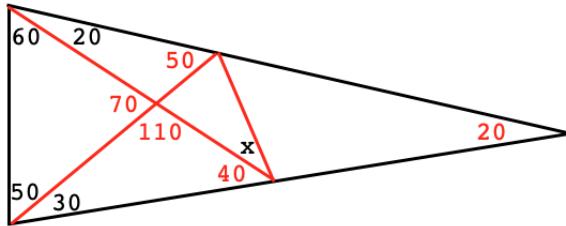
Chapter 25

Langley

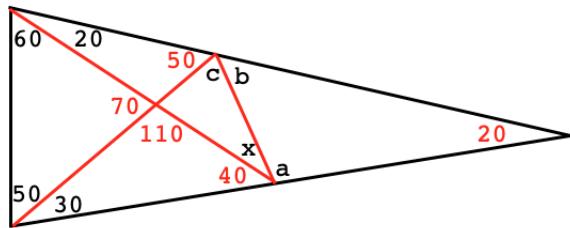
Here is a problem that looks like it will be easy, and then turns out, famously, to be hard. Let's take a look.



Using theorems about supplementary and vertical angles, and triangle sums, we can fill in some of the values:



However, we're not any closer to x . One idea is to use algebra:



So now we can use the triangle sum of angles theorem to write some equations:

$$a + x = 140$$

$$a + b = 160$$

$$b + c = 130$$

$$c + x = 110$$

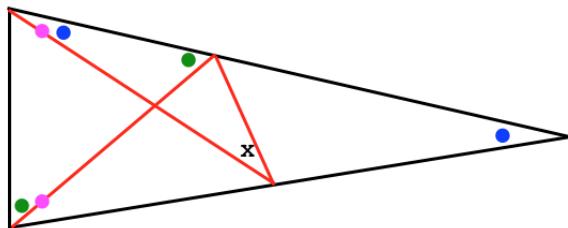
Four equations in four unknowns. That could work. However, subtract the first from the last and the second from the third to obtain:

$$a - c = 30$$

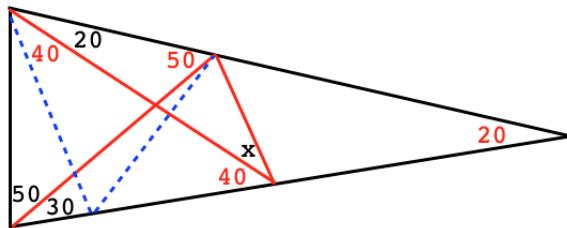
$$a - c = 30$$

In linear algebra, we call that degenerate. The system is not independent and we cannot solve it.

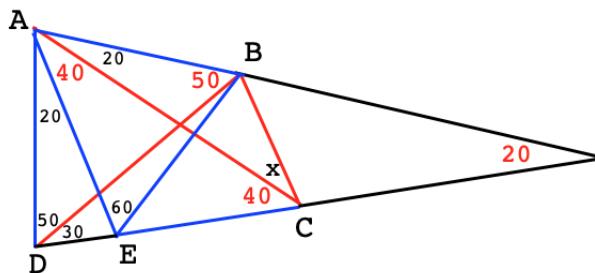
However, we notice some isosceles triangles. I found three:



This will bring us to the "bright idea."



We pick another point on one of the sides and draw the two line segments so that the angle labeled at the top is 40° .



So then we can do some geometry. First of all, the total angle at vertex A is 80 . Therefore, the angle DAE is 20 . Since the angle at vertex D is 80 , $\triangle ADE$ is isosceles and $AD = AE$.

We already knew that $\triangle ABD$ is isosceles, so $AD = AB$. Thus, $AE = AB$. So now $\triangle ABE$ is isosceles, and since angle BAE is 60 , the triangle is also equilateral. So $BE = AE = AB$.

By equal base angles $\triangle ACE$ is isosceles, so $AE = EC$.

By adding up angles, we can find that angle BEC equals 40 . Since $\triangle BEC$ is isosceles, $x + 40 = (1/2)140 = 70$ and then $x = 30$.

There is a lot more.

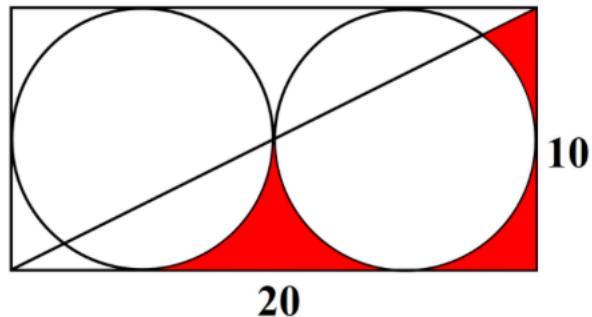
https://en.wikipedia.org/wiki/Langley%27s_Adventitious_Angles

Chapter 26

Circular arch

I found a hard geometry problem on the web:

HARD: Find the total area of the red spots.



We know it's hard because it says so!

I liked it particularly because there are several different ways to calculate the answer using basic geometry and trigonometry, plus standard integration as well as polar integration. I got the same answer each time, fortunately.

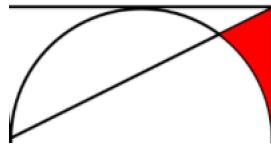
In this book we just look at the geometry.

As a first step, observe that the problem has been made artificially complicated by using these particular values for the side lengths. If both dimensions are scaled down by a factor of 5, then we obtain a rectangle with side lengths 4 and 2, and the two circles become unit circles.

We must just remember to re-scale to obtain the final answer, multiplying the final area by 25.

The area of any arched corner segment is pretty easy, since 4 of them put together are equal to the difference between the area of a square with side length 2 and a unit circle: $4 - \pi$, so each one is $1 - \pi/4$.

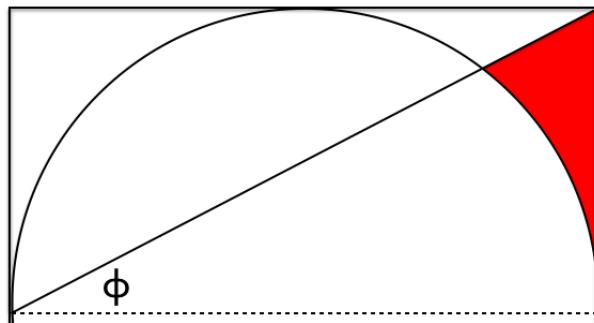
The real difficulty is the upper right-hand corner. We ignore the above and just concentrate on this part of the problem.



One of the arches is divided into two pieces, and we are supposed to only count the red part of the arch.

The basic right triangle that we see repeated in these images has side lengths in the ratio $1 : 2$. Its area is just 1, and the smaller angle is

$$\phi = \tan^{-1} 0.5 \approx 0.4636 \text{ rad } \approx 26.565 \text{ deg}$$



That's not a nice round number, but OK.

My first thought was to calculate the area cut off by the chord of a circle, called a "circular segment". Then we could calculate the white part of the divided arch:

triangle – segment – arch

$$1 - \text{segment} - (1 - \pi/4)$$

and subtract that from one whole arch to get the red part.

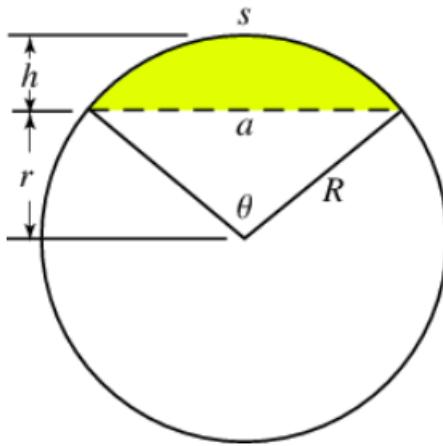
$$(1 - \pi/4) - [1 - \text{segment} - (1 - \pi/4)]$$

$$1 - \frac{\pi}{2} + \text{segment}$$

We will use this result at the end for our final answer.

There is an easier way, which we find by exploring this direction just a little further.

A circular segment is like a polar cap, but in two dimensions.



<http://mathworld.wolfram.com/CircularSegment.html>

We carefully distinguish between the circular segment, in yellow, and the circular sector, which is the area of that slice of the circular pie swept out by the angle θ .

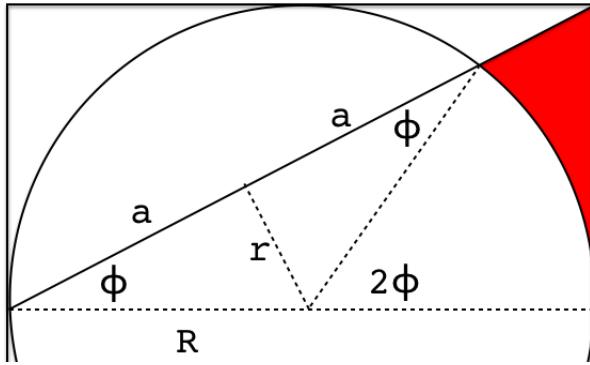
The area of the circular sector with central angle θ is the fraction of the total circular angle, times the area of a unit circle. The result is just half the angle.

$$\frac{\theta}{2\pi} \cdot \pi = \frac{\theta}{2}$$

For the actual calculation of a circular segment, we would need not only the angle θ , but also r and a , which we would need to derive from θ by applying the Pythagorean theorem and/or trigonometry. It can be done! However, we see a better way.

isosceles triangles

The key idea is to draw another radius on our diagram, and realize that θ is the apex angle of an isosceles triangle. The two sides are both radii of our circle, and so are equal to each other! Therefore $\theta = \pi - 2\phi$.



We recall that this is a basic theorem from the geometry of circles: the central angle subtending a given arc is twice the measure of the angle on the periphery.

Now we're ready.

The area of the circular segment with a central angle θ is

$$A = \frac{\theta}{2\pi} \cdot \pi R^2$$

This is a unit circle so $R = 1$ and then

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

Since the angle we're talking about is actually 2ϕ we obtain finally

$$A = \phi$$

That's for the area of the circular segment.

For the triangles, we see that r is the bisector of an isosceles triangle. Therefore

$$\frac{r}{R} = r = \sin \phi$$

and

$$\frac{a}{R} = a = \cos \phi$$

So the area of the two triangles is

$$\sin \phi \cos \phi$$

and the area of the red part of the arch is then the area of the large right triangle ($1/2 \cdot 1 \cdot 2 = 1$) minus the two areas we just calculated:

$$1 - \phi - \sin \phi \cos \phi$$

One could change the angle ϕ and this result would still be valid, except that the area of the large triangle would change.

The base would still be 2, and the height would be $2 \tan \phi$, so the area would be $2 \tan \phi$ rather than 1.

using the tangent

This particular angle has tangent 1/2, which gives an additional simplification:

$$\frac{1}{2} = \frac{\sin \phi}{\cos \phi}$$

$$2 \sin \phi = \cos \phi$$

$$2 \sin^2 \phi = \sin \phi \cos \phi$$

We can do even better than that.

$$2 \sin \phi = \cos \phi$$

$$4 \sin^2 \phi = \cos^2 \phi$$

so

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

$$5 \sin^2 \phi = 1$$

$$2 \sin^2 \phi = \frac{2}{5}$$

Our final answer for this part is

$$1 - \phi - \frac{2}{5}$$

To get the answer required by the original problem statement, we must add the area of three arches, $3 \cdot (1 - \pi/4)$, and remember to scale the answer back up by a factor of 25.

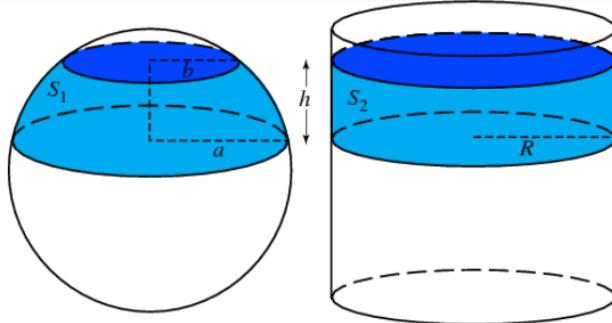
Chapter 27

Hatbox theorem

A famous result of Archimedes, among many others, is called the hat-box theorem, which states that the surface area of a sphere is equal to the lateral surface area of a cylinder which just encloses it, as we had above. (Lateral area does not count the end pieces).

For a sphere and cylinder of radius R , the cylinder has surface area of the circumference $2\pi R$ times the height $2R$ for a total of $4\pi R^2$.

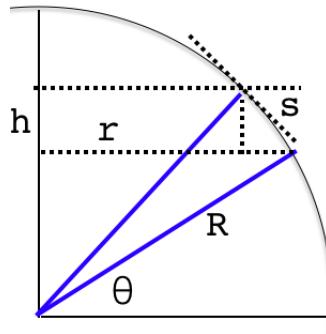
Archimedes showed that this is true not just for the whole, but for any slice or section through the sphere. That's pretty amazing.



Let's sketch a geometric proof briefly.

First, consider a thin strip of surface area extending around the sphere on a great circle (such as the equator). The surface area will be the circumference times the width of the belt, or $2\pi R \times h$.

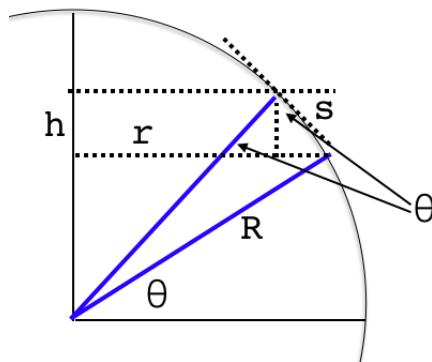
Second, this is true even for a belt that is not at the equator. Consider this figure:



What is the surface area contained between two horizontal cuts of the sphere along the dotted lines? Such a strip is called a "spherical belt". If the slice is very thin, then the circumference at the top and bottom of the slice will be approximately the same, with radius $r = R \cos \theta$.

To get the surface area, we must multiply the circumference $2\pi r$ by the width of the belt. The width is not the height h but s (called the slant height), because of the tilt of the surface.

For a very thin slice, the angle θ won't change much in going from the first blue radius R to the second one. Seeing this, it is then not hard to work out that the angle between s and h in the right triangle containing them both is equal to θ



so

$$\cos \theta = \frac{h}{s}$$

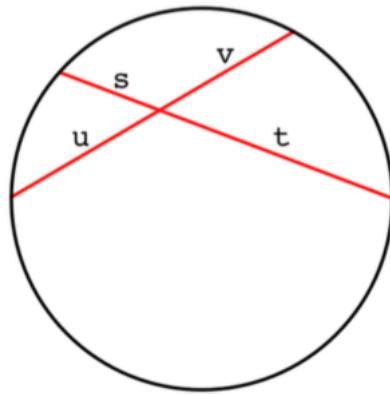
The area is

$$a = 2\pi rs = 2\pi R \cos \theta \frac{h}{\cos \theta} = 2\pi Rh$$

The same as before.

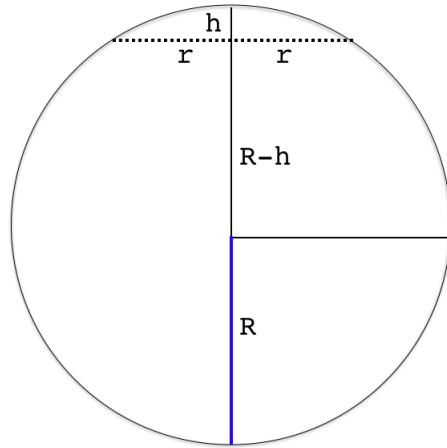
Third, consider the "belt" at the very top of the sphere. This region is more of a "spherical cap", like a contact lens or one of the poles of earth.

We recall a famous result concerning chords of a circle.



$$st = uv$$

This relationship holds for any two chords of the circle. In particular, it holds for the chord consisting of $r + r$ and the chord consisting of $h + (2R - h)$.



By the above theorem

$$r^2 = h(2R - h) = 2Rh - h^2 \approx 2Rh$$

Since h is small, we can neglect a factor of h^2 .

The area of the circle is

$$a = \pi r^2 = 2\pi Rh$$

so we have the same rule as above.

Therefore, for *every* belt of height h , the area is $2\pi Rh$.

In summing up the contributions from each belt of width h_i

$$A = \sum a = \sum 2\pi Rh_i = 2\pi R \sum h_i$$

But $\sum h_i$ is just equal to R so we have

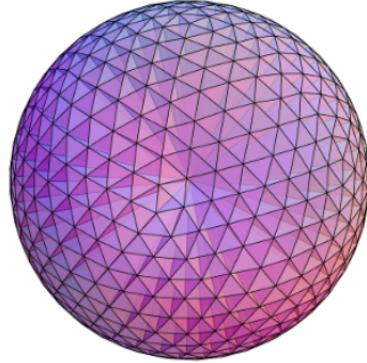
$$A = 2\pi R^2$$

The surface area of a hemisphere of radius R is twice the area of its great circle. Thus, the total area of the sphere is $4\pi R^2$.

geometric proof

Here is a geometric proof that assumes we know the volume of the sphere is

$$V = \frac{4}{3} \pi R^3$$



Divide the whole sphere up into triangular prisms. Each one has volume $dV = 1/3 R dA$. So for the whole thing the volume =

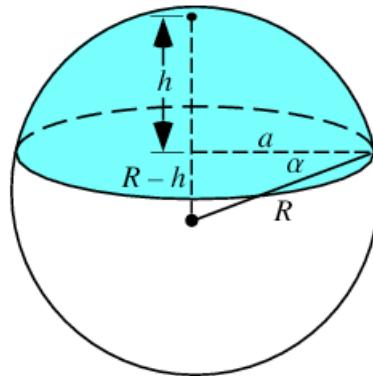
$$\frac{4}{3} \pi R^3 = \frac{1}{3} R A$$

$$A = 4\pi R^2$$

Chapter 28

Spherical cap

Here is a figure from Wolfram for a spherical cap. We are interested in formulas for the area and volume of the solid obtained by slicing through a sphere, where the height of the cap that is produced is h , and the distance of closest approach to the center of the sphere is $R - h$.



geometry

If we start from the equator, and think about a thin belt going around the sphere, the belt has length equal to the circumference $2\pi R$ and width h , and thus area S :

$$S = 2\pi Rh$$

We believe this should be the formula for the surface area of a belt of width h , at least near the equator. In the figure, this width is labeled as $R - h$, because we

are more interested in the cap. Thus, for the calculation below, this area will be $2\pi R(R - h)$.

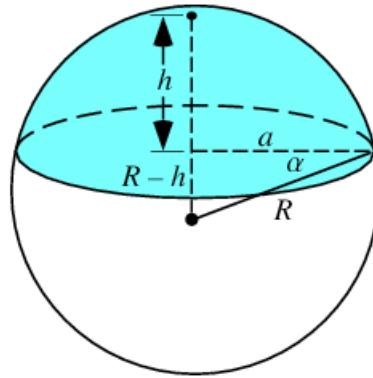
Consider that the total surface area of the hemisphere is $2\pi R^2$ so the area of the cap is the difference

$$S = 2\pi R^2 - 2\pi R(R - h) = 2\pi Rh$$

That's a surprising result, that the area of the cap depends only on R and its width (here called h). At least, that is certainly true in the limit as the width of the belt at the equator is very small.

polar cap

Furthermore, if we look in the figure at the right triangle with h and a as the sides, we can draw the hypotenuse of that triangle and call it r (it's not actually labeled in the figure).



It is sometimes called the slant height. We calculate

$$a^2 = R^2 - (R - h)^2 = 2Rh - h^2$$

$$r^2 = a^2 + h^2 = 2Rh - h^2 + h^2 = 2Rh$$

Now think about a very small spherical cap, then it would be almost flat, a circle, and its radius would be r and area

$$S = \pi r^2$$

But $r^2 = 2Rh$, so again we have the same formula for the surface area of a small cap and a belt near the equator!

General case

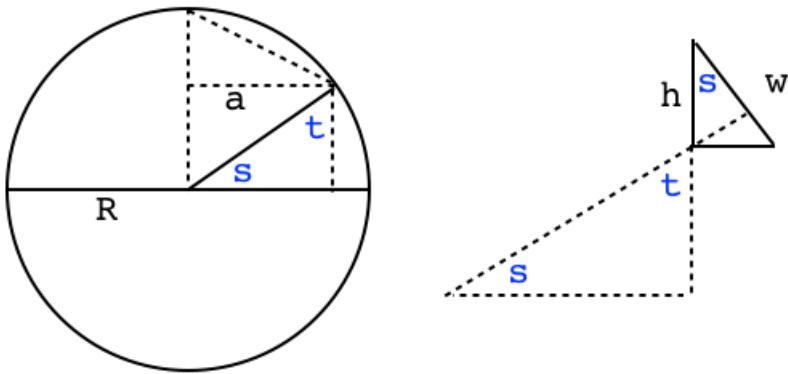
Consider a belt of width h at a position somewhere in the temperate latitudes of the sphere, not close to either the pole or the equator.

We use a thin belt, so that going north toward the pole the surface of the sphere is approximately flat. As before, h is the width of the projection of the belt on the z -axis.

The width w of the belt on the surface is larger than h , because w is not vertical but tilted toward the z -axis. And since the surface is flat, this angle with the z -axis is constant over the width of the belt.

Draw a ray from the center of the sphere to the point where the belt is. The ray makes an angle s with the xy -plane, at the center of the sphere. The radius a at the position of the belt is smaller than R by a factor of $\cos s$.

$$a = R \cos s$$



The tangent to the sphere at this point (namely, w) is perpendicular to the ray. In the right panel, we see a small triangle on the surface of the sphere with sides w and h .

All three of the triangles shown in the right panel are right triangles, with complementary angles s and t (not all of them are labeled). Can you see that the angle between h and w is s ?

Therefore, the slant height w of the belt is larger than h by the factor of $\cos s$.

$$h = w \cos s$$

So the true area is

$$2\pi a w = 2\pi R \cos s \frac{h}{\cos s} = 2\pi Rh$$

The cosine of the angle comes in twice, and these factors cancel.

The formula $2\pi Rh$ is correct everywhere.

Chapter 29

Pappus

Pappus' centroid theorem is actually a pair of theorems about solids of revolution, where a curve C is revolved around a central axis. The two theorems relate to the surface area and volume.

There is a nice article about it at Mathworld:

<http://mathworld.wolfram.com/PappussCentroidTheorem.html>

Statements

The first theorem states that the surface area SA is the product of the arc length s of the curve C times the distance d traveled by the geometric centroid of C .

The example in the wikipedia article is a torus of minor radius r and major radius R . Then C is the circle of radius r , the centroid of the curve is its center, and this point moves around a circle of radius R the distance $2\pi R$.

The first term is the circumference C of the curve (the small circle) and the total is

$$SA = 2\pi r \cdot 2\pi R = 4\pi^2 rR$$

One puzzling thing is that we are used to taking account of the slant height in thinking about surface area (though not volume), but the curvature of the surface doesn't seem to be an issue here.

The second theorem depends on the area enclosed between the curve and the axis of rotation.

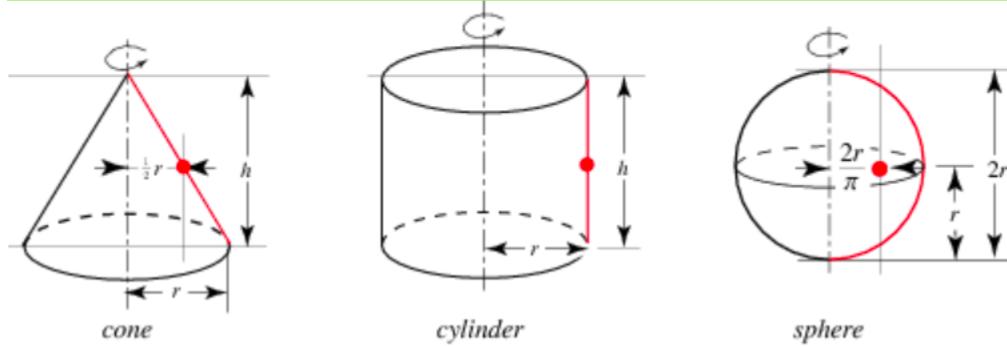
The volume of the solid is the product of this area times the distance d traveled by the geometric centroid of the area (not the curve). For the torus or donut

$$V = \pi r^2 \cdot 2\pi R = 2\pi^2 Rr^2$$

The main trick here for both theorems is to actually find the geometric centroids. We'll work the two easy cases first: cylinder and cone.

Cylinder surface area

Here is a picture from Wolfram. Our notation is slightly different.



For the cylinder, revolve a parallel line segment around the y -axis. The curve has length H .

The centroid of the parallel line segment (the average distance of each point on the curve from the y -axis) is just the radial distance R , since all the points are the same distance. In addition, the centroid is also halfway along the curve at $H/2$.

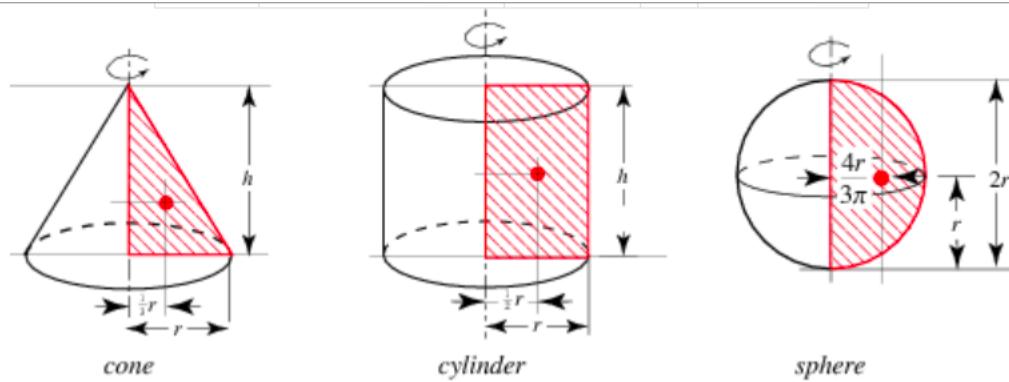
The distance this point travels is $2\pi R$.

The surface area is the product of the arc length H and the distance traveled by the centroid during the revolution

$$SA = H \cdot 2\pi R = 2\pi RH$$

The classical way to obtain a formula for the surface area of the cylinder is to imagine cutting along the length of it, forming a rectangle with width H and length $2\pi R$, which gives the same result.

Cylinder volume



To find the volume of the cylinder, we need to consider the area enclosed by the curve and the y -axis (called a lamina).

For the cylinder, the centroid of the lamina is at $R/2$ and its area is RH . Multiply the area times the distance traveled by the centroid

$$V = RH \cdot 2\pi \frac{R}{2} = \pi HR^2$$

Cone area

For the cone, we revolve an inclined line segment around the y -axis, with one end lying on the axis and the other at the radius R .

The centroid of this line segment lies at a distance $R/2$ from the y -axis. The distance it travels during the rotation is then πR .

The length of the curve is the slant height s . Multiplying to get the surface area

$$A = \pi R s$$

The classical way to obtain this formula for the surface area of a cone (which we saw in a previous chapter) is to imagine cutting along the slant of the cone to obtain a

sector of a circle. The circle has radius s and circumference $2\pi s$ and area πs^2 . We take the ratio of the outer perimeter of the sector to the whole circumference, times the whole area

$$\frac{2\pi R}{2\pi s} \pi s^2 = \pi R s$$

Cone volume

For the volume of the cone, we need to look at the triangle formed from the inclined line segment. Its area is $Rh/2$. Now, what is its geometric centroid?

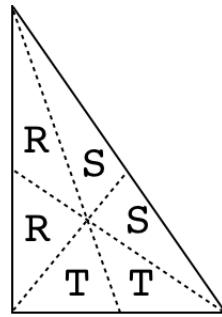
To begin with, I will just use the available result from wikipedia, which is that the centroid of a right triangle is $1/3$ of the distance along each side away from the right angle, i.e. $R/3$. The distance it travels during the rotation is then $2\pi R/3$.

The area of the triangle is $(1/2)RH$ and so the volume is

$$V = \frac{1}{2}RH \cdot 2\pi \frac{R}{3} = \frac{1}{3}\pi R^2 H$$

Centroid of the cone's lamina

For the triangle, we compute the centroid by a geometric argument. The first part of the following holds for any triangle, but I've drawn a right triangle because that's what we've got in the problem (for the cone).



We draw lines from each vertex to the midpoint of the opposite side. The three lines cross at a single point, the centroid. (We looked at the proof of this in Ceva's Theorem). It is easy to see that the areas of the small triangles with the same letter are equal.

For example, both triangles T have the same base (because we drew the median), and the same height. For the same reason

$$R + R + T = S + S + T$$

$$R + R = S + S$$

That is

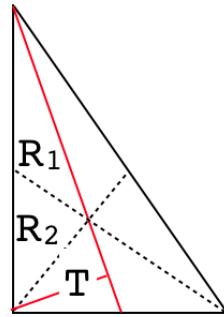
$$R = S$$

As well

$$R = S = T$$

The extension to T follows because the problem is symmetrical.

Now consider the median shown in red in the figure below and the altitude drawn to it.



Both the triangle labeled T and the triangle formed from $R_1 + R_2$ have the same height, namely the altitude drawn in red. But the area of $R_1 + R_2$ together is twice that of T .

Therefore the length of the base of $R_1 + R_2$ (along the median shown in red) must be twice that for the triangle labeled T . The centroid — the point where the lines meet — lies at $2/3$ of the distance from the vertex to the opposing side or $1/3$ of the way up from the bottom

Since we have a right triangle, then by similar triangles, both the x -coordinate and the y -coordinate of the centroid are at $1/3$.

This result can be produced by several different arguments. By looking at Ceva's Theorem geometrically, by using vectors, or by calculus (in the chapter on average value in that book), and now again here. It is not only reassuring to find the same answer each time, but positively required.

There is yet another proof that comes from Euler's construction of the three points on the same line: orthocenter, centroid and circumcenter, and showing that the distance from centroid to orthocenter is twice that from centroid to circumcenter.

The rest of the discussion of Pappas theorems is best with calculus. We defer that to a later book.

Part VIII

Addendum

Chapter 30

More on sum of angles

Here are several other derivations that I've come across over the years. You only need what we've given above, but they are interesting to work through and give practice in dealing with trig functions.

calculus

All you need to know for this is that the derivative of sine is cosine and the derivative of the cosine is minus the sine. If you don't know about derivatives yet, just skip this example.

Suppose we know one of the formulas already, say

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

Treat t as a constant and take the derivative with respect to s . The left-hand side is

$$\frac{d}{ds} \cos s + t = -\sin s + t$$

whlle the right-hand side is:

$$\frac{d}{ds} (\cos s \cos t - \sin s \sin t) = -\sin s \cos t - \cos s \sin t$$

But these are equal! Multiplying by -1 , we have that

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

Going the other way works as well, as does treating t as the variable.

vector rotation

I want to take a moment to introduce a concept that gives another simple proof of the sum of angles theorems. We don't have time to cover it in the detail it really deserves, that's for later.

Start with the idea of a *vector* from the origin to a point. Vectors have length — the vectors depicted are unit vectors with length 1. They also have a direction. The unit vector along the x -axis can be represented as the pair of x, y -coordinates $(1, 0)$.

Let us rotate the vector in the counter-clockwise direction by an angle ϕ . The new coordinates of the head of the vector (marked with the arrow) are $(\cos \phi, \sin \phi)$. We have

$$1, 0 \rightarrow \cos \phi, \sin \phi$$

for a vector x units long we have

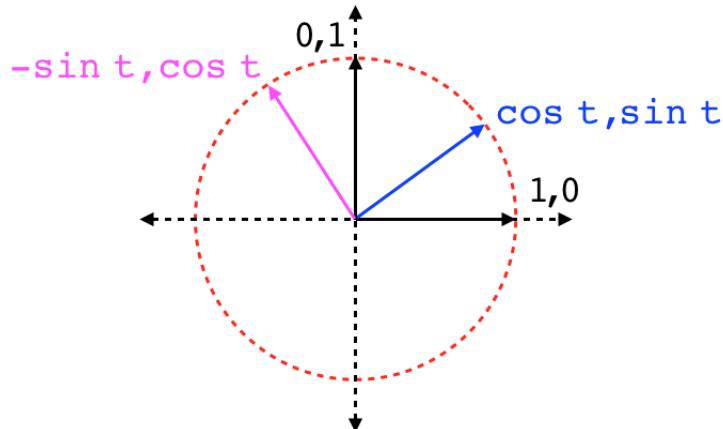
$$x, 0 \rightarrow x \cos \phi, x \sin \phi$$

By similar reasoning

$$0, 1 \rightarrow -\sin \phi, \cos \phi$$

$$0, y \rightarrow -y \sin \phi, y \cos \phi$$

We get a minus sign because the rotated unit vector that started pointing straight up is now in the second quadrant, the x -component is negative.



For a general vector (x, y) the resulting vector is (x', y') and the components are:

$$x' = x \cos \phi - y \sin \phi$$

$$y' = x \sin \phi + y \cos \phi$$

Now, suppose we rotate a second time, by an angle θ . Just add a prime for each and substitute θ for ϕ .

$$x'' = x' \cos \theta - y' \sin \theta$$

$$y'' = x' \sin \theta + y' \cos \theta$$

Write x'' in terms of the original x and y :

$$\begin{aligned} x'' &= x' \cos \theta - y' \sin \theta \\ &= (x \cos \phi - y \sin \phi) \cos \theta - (x \sin \phi + y \cos \phi) \sin \theta \\ &= x \cos \phi \cos \theta - y \sin \phi \cos \theta - x \sin \phi \sin \theta - y \cos \phi \sin \theta \\ &= x [\cos \phi \cos \theta - \sin \phi \sin \theta] - y [\sin \phi \cos \theta + \cos \phi \sin \theta] \end{aligned}$$

The key idea is that we must get the same result if we turn through both angles at once:

$$x'' = x(\cos \phi + \theta) - y(\sin \phi + \theta)$$

The cofactors of x and y must separately be equal:

$$\cos \phi + \theta = \cos \phi \cos \theta - \sin \phi \sin \theta$$

$$\sin \phi + \theta = \sin \phi \cos \theta + \cos \phi \sin \theta$$

Those are our formulas!

If you know about matrix multiplication, you can write

$$\begin{bmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{bmatrix}$$

and then

$$\begin{bmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{bmatrix} \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} = \begin{bmatrix} \cos s + t & \sin s + t \\ -\sin s + t & \cos s + t \end{bmatrix}$$

The simple multiplication rule will recover our formulas.

Chapter 31

Pythagorean triples

In a previous chapter we derived what are called the double-angle formulas:

$$\sin 2s = 2 \sin s \cos s$$

$$\cos 2s = \cos^2 s - \sin^2 s$$

We will manipulate these to find expressions in terms of the same variable, using the following identity:

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

$$\tan^2 \theta + 1 = \frac{1}{\cos^2 \theta}$$

sine

$$\begin{aligned}\sin 2s &= 2 \sin s \cos s \\&= 2 \frac{\sin s}{\cos s} \cos^2 s \\&= 2 \tan s \frac{1}{1 + \tan^2 s}\end{aligned}$$

Let $a = \tan s$, then

$$\sin 2s = \frac{2a}{1 + a^2}$$

cosine

$$\begin{aligned}\cos 2s &= \cos^2 s - \sin^2 s \\&= \left[\frac{\cos^2 s}{\cos^2 s} - \frac{\sin^2 s}{\cos^2 s} \right] \cos^2 s \\&= \left[\frac{1 - \tan^2 s}{1 + \tan^2 s} \right]\end{aligned}$$

so

$$\cos 2s = \frac{1 - a^2}{1 + a^2}$$

triples

In general, a can be anything. But if a is a rational number, then we can obtain the corresponding sides of a right triangle with rational lengths as well.

The sides are: $2a, 1 - a^2$ with the hypotenuse:

$$\begin{aligned}&\sqrt{4a^2 + (1 - 2a^2 + a^4)} \\&\sqrt{1 + 2a^2 + a^4} \\&= 1 + a^2\end{aligned}$$

Suppose $a = \frac{2}{3}$. Then, we have side lengths: $\frac{4}{3} = \frac{12}{9}, \frac{5}{9}$, and $\frac{13}{9}$, which can be converted to integers: 12, 5, 13.

In general, if $\tan s = p/q$ then the sides are

$$\frac{2p}{q}, \quad 1 - \frac{p^2}{q^2}, \quad 1 + \frac{p^2}{q^2}$$

which as integers will be

$$2pq, \quad q^2 - p^2, \quad q^2 + p^2$$

This formula was found by Euclid.

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

If p and q are two odd integers the sum and difference of squares is even so we can write

$$pq, \quad \frac{q^2 - p^2}{2}, \quad \frac{q^2 + p^2}{2}$$

As another example, let $q = 5$, $p = 3$, and we have 15, 8, 17, another triple.

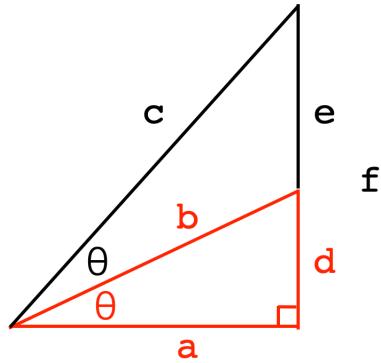
Chapter 32

Archimedes and pi

We're going to follow a page that presents Archimedes' method for the approximation of π .

<https://itech.fgcu.edu/faculty/clindsey/mhf4404/archimedes/archimedes.html>

To begin with, let's review some ideas related to angle bisection. Recall that if we have an angle bisector in a right triangle



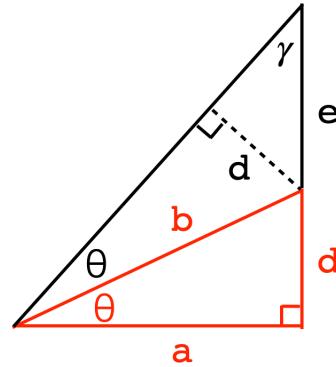
the theorem says that

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

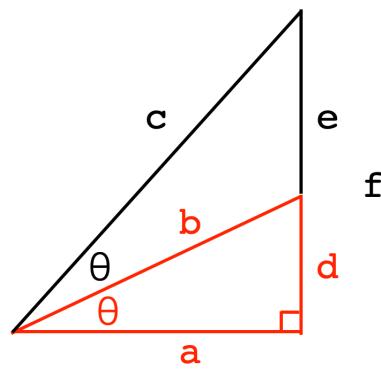
The proof (which we showed in the the Geometry book) involves drawing the altitude

of the top triangle, forming two congruent triangles and one more that is similar to the original.

For congruency, we have two (that is, three) angles the same plus a shared side, b . That accounts for the label of d on the dotted line.



The small triangle at the top is similar to the one we started with, by complementary angles in a right triangle. Let us now re-label the hypotenuse of the original.



By similar triangles, we have (adjacent over hypotenuse):

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

This is rearranged to give

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

□

There's more. Add 1 to both sides:

$$\begin{aligned}\frac{a+c}{c} &= \frac{d+e}{e} \\ \frac{a+c}{d+e} &= \frac{c}{e}\end{aligned}$$

The denominator on the left-hand side is just f , and the right-hand side is equal to a/d so

$$\begin{aligned}\frac{a+c}{f} &= \frac{a}{d} \\ \frac{a}{f} + \frac{c}{f} &= \frac{a}{d}\end{aligned}$$

□

This is the central relationship we will use. If we simply add the cotangent and cosecant of the original angle, we obtain the cotangent of the half-angle, a/d .

To get the cosecant (b/d), use the Pythagorean theorem.

$$\begin{aligned}a^2 + d^2 &= b^2 \\ \frac{a^2}{d^2} + 1 &= \frac{b^2}{d^2}\end{aligned}$$

Square the cotangent, add 1, and take the square root. That's b/d , the cosecant, or $1/\sin$.

Of course, Archimedes does not use the language of trigonometry. He views these numbers simply as ratios (*logos*).

overview

There are three steps which we will repeat four times.

- At each round, we start with the values for the cosecant and cotangent (to start with, c/f and a/f). In keeping with Greek tradition, these are ratios of whole numbers.

- Add them together, obtaining $(c + a)/f$, which is equal to a/d . This gives us the *cotangent* of the half-angle.
- Use the Pythagorean theorem to obtain the *cosecant* of the half-angle, b/d .

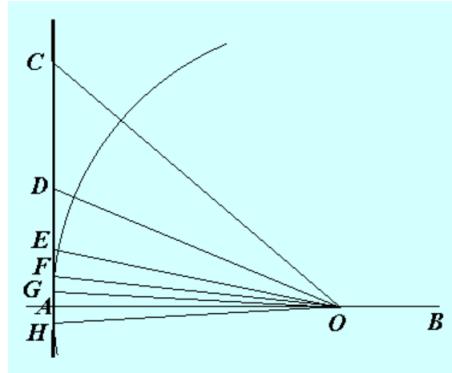
In what follows first (part C) we compute the circumference or perimeter of a circumscribed polygon; after that we'll compute the inscribed perimeter in part I.

Because the first one is circumscribed, we're finding an *upper* bound for the value of π .

C1

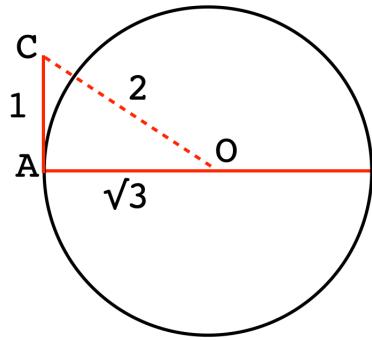
Draw a circle with radius OA and tangent AC . (A is hard to see, it lies between G and H near the bottom of the figure).

Let the central $\angle AOC$ be one-third of a right angle. Thus $\angle AOC$ is 30° .



The figure appears to have been compressed in width. The angle bisectors don't look right and the original angle looks more like 45 than 30 . We'll use it anyway since it comes from the reference.

Recall that if we draw an altitude in an equilateral triangle, the side at the base is bisected, which means that the sides of the two right triangles are in the ratio $1 : \sqrt{3} : 2$. The adjacent side for the smaller angle (30°) is $\sqrt{3}$ and the opposite side is 1, so the cotangent (ratio $OA : AC$) is $\sqrt{3}$, the square root of three.



$265/153$ is a (very good) approximation to $\sqrt{3}$, just slightly smaller than the true value. We explore how Archimedes may have done this calculation elsewhere.

We shouldn't get ahead of ourselves, but you might wonder how good an approximation for π we have to start with.

The central angle is 30° . The length AC is one-half of a side from a regular hexagon. So the total perimeter is 12 times that. If we invert the cotangent and multiply by 12, we'll have the ratio of the circumference to the radius. Since π is the ratio to the diameter

$$\begin{aligned}\pi &= \frac{C}{d} = \frac{C}{2r} \\ &= \frac{1}{2} \cdot \frac{C}{r}\end{aligned}$$

we multiply by 6 rather than 12.

The preliminary estimate is $6 \cdot 1/\sqrt{3} = 3.4641$.

At each step we will bisect the central angle. The number of sides will double, so the tangent will be multiplied by 12, 24, 48 and finally by 96.

Archimedes' calculation

In what follows we list the claim first:

- $OA : AC > 265 : 153$

Followed by the

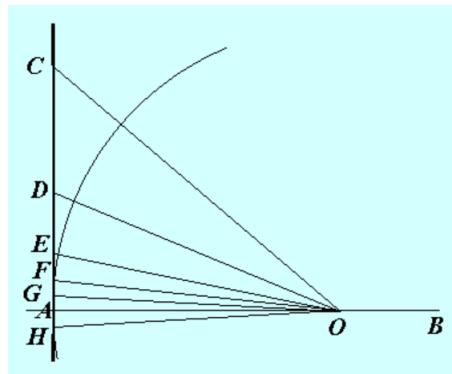
Proof.

This is just Archimedes' estimate for $\sqrt{3}$, slightly less than the true value. Next

$$\circ OC : AC = 306 : 153$$

The cosecant is equal to 2. The denominator of the ratio has been chosen to match the previous one.

Now draw the angle bisector OD .



The new cotangent ($OA : AD$) of the bisected angle is the sum of the original cotangent and cosecant, as discussed above.

$$\circ OA : AD > 571 : 153$$

We have just added the numerators for the first two ratios above, leaving the result over the common denominator.

$$\circ OD : AD > 591 - 1/8 : 153$$

We want the new cosecant. Square the numerator of the previous result and add the square of 153, then take the square root.

So

$$571^2 = 326041$$

$$153^2 = 23409$$

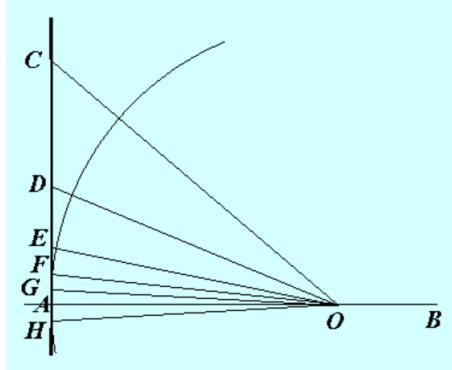
$$326041 + 23409 = 349450$$

A modern computation of the square root of 349450 gives about 591.143 in decimal.

Archimedes approximates the result as $591 \frac{1}{8}$. Place that over 153 to obtain the ratio shown above.

C2

Now draw the angle bisector OE .



- From above, we have that $OA : AD > 571 : 153$ and $OD : AD > 591 \frac{1}{8} : 153$.
- $OA : AE > 1162 \frac{1}{8} : 153$

This calculation invokes the angle bisector corollary again. Rather than repeat the derivation, just add the inputs:

$$591 \frac{1}{8} : 153 + 571 : 153$$

which adds to give the result above, $1162 \frac{1}{8} : 153$.

- $OE : AE > 1172 \frac{1}{8} : 153$

Use the Pythagorean theorem to write:

$$OE^2 = AE^2 + OA^2$$

$$\frac{OE^2}{AE^2} = \frac{OA^2}{AE^2} + 1$$

$$1162\frac{1}{8}^2 = 1350534\frac{1}{2}$$

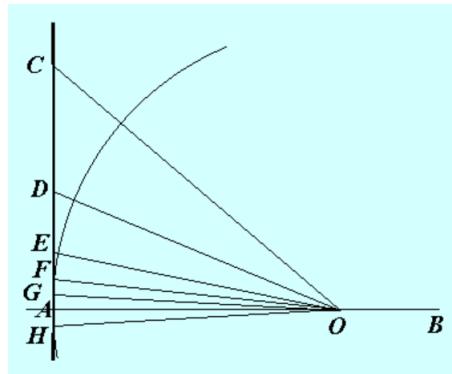
$$153^2 = 23409$$

$$1350534\frac{1}{2} + 23409 = 1373943\frac{1}{2}$$

A modern computation of the square root gives 1172.15.

C3

Now draw the angle bisector OF .



- From above, we have that $OA : AE > 1162 \frac{1}{8} : 153$ and $OE : AE > 1172 \frac{1}{8} : 153$.
- $OA : AF > 2334 \frac{1}{4} : 153$

This calculation invokes the angle bisector corollary again.

$$\frac{OA}{FA} = \frac{OE}{EA} + \frac{OA}{EA}$$

$$1162 \frac{1}{8} : 153 + 1172 \frac{1}{8} : 153$$

which adds to give the result above.

- $OF : FA > 2339 \frac{1}{4} : 153$

Use the Pythagorean theorem to write:

$$\frac{OF^2}{FA^2} = \frac{OA^2}{FA^2} + 1$$

$$2334 \frac{1}{4}^2 = 5448723 \frac{1}{8}$$

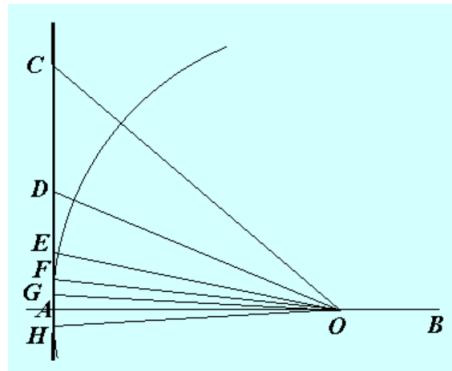
$$153^2 = 23409$$

$$5448723 \frac{1}{8} + 23409 = 5472132 \frac{1}{8}$$

The square root is $2339 \frac{1}{4}$.

C4

Now draw the angle bisector OG .



- From above, we have that $OA : FA > 2334 \frac{1}{4} : 153$ and $OF : FA > 2339 \frac{1}{4} : 153$

Add

- $OA : AG > 4673 \frac{1}{2} : 153$

We're almost done. We do not need to compute the cosecant at this last step

The original multiplier was 6. There is an additional factor for the four "halvings" of $2^4 = 16$. Hence we obtain

$$153 \times 96 = 14688$$

and then invert to get the ratio of the circumference to the diameter:

$$\frac{14688}{4673 \frac{1}{2}} = 3 + \frac{667 \frac{1}{2}}{4673 \frac{1}{2}}$$

The fraction is just less than $1/7$.

$1/7 = 0.142857$, while $667 \frac{1}{2}/4673 \frac{1}{2} = 0.142827$.

We conclude that $\pi < 3 \frac{1}{7}$.

Check our progress

Here are calculations for the upper bound at the beginning and then after each of four steps:

```
>>> 6 * 153/265.0
3.4641509433962265
```

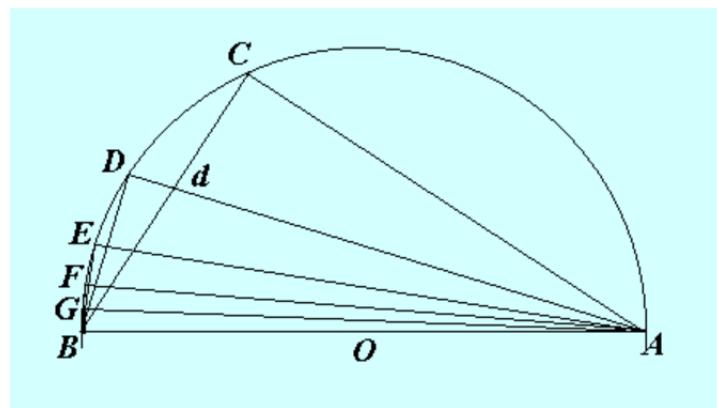
```

>>> 12 * 153/571.0
3.2154115586690017
>>> 24 * 153/1162.125
3.15972894482091
>>> 48 * 153/2334.25
3.1461925672057407
>>> 96 * 153/4673.5
3.1428265753717772
>>>
>>> 1/7.0
0.14285714285714285

```

Part I

Here is the diagram for an inscribed polygon.



As before $\triangle ABC$ is a $30 - 60 - 90$ right triangle, but now it lies inside the circle.

- $AC : BC < 1351 : 780$.

AC/BC is the cotangent of 30° .

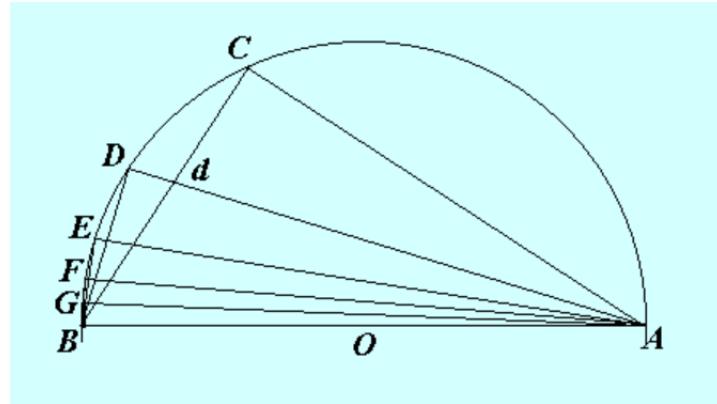
The ratio $1351/780$ is an approximation for $\sqrt{3}$. It is an even better approximation than the previous one, and also important, it is just slightly *more* than the true value, whereas $265/153$ was slightly less.

Next, we will bisect $\angle CAB$ to give $\angle DAB$ and use the same relationship that we had before: the sum of the cotangent plus the cosecant of the original angle is equal to the cotangent of the half angle.

On one hand, this is just the bisector theorem (which gives the result for $\angle CAD$), we need to show that the cotangent of $\angle CAD$ is equal to the cotangent of $\angle DAB$ ($= \angle DAB$). But we bisected the original angle, so of course they are equal as given.

Nevertheless, Archimedes can't do it that way.

We need, essentially, a different proof of the bisector theorem that applies to $\triangle ABD$.



I1

Let AD bisect the angle, and then join BD .

- $\angle BAD = \angle DAC = \angle DBd$.

The first statement just restates the construction as an angle bisector. The second follows from the fact that $\angle BDd$ is a right angle, which establishes $\triangle BDd \sim \triangle AdC$.

- $AD : DB < 2911 : 780$

According to our source, using the similar triangles above, write three ratios:

$$\begin{aligned} AD : BD &= BD : Dd = AB : Bd \\ &= (AB + AC) : (Bd + Cd) \\ &= (AB + AC) : BC \end{aligned}$$

*HT
GT*

or $(BA + AC) : BC = AD : DB$.

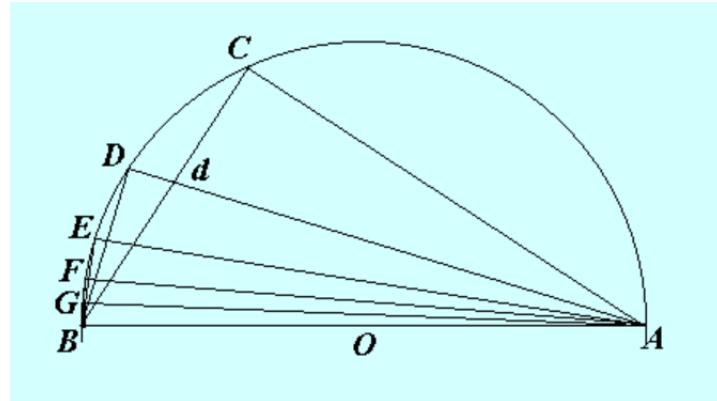
$$AD : BD = BD : Dd = AB : Bd$$

This seems to be an error. The first two are cotangents, while the third is the ratio of a hypotenuse to the short side in a triangle that is not a right triangle!

However, conclusion

$$(AB + AC) : BC = AD : DB$$

is easily proved as follows:



Proof.

We have that $\triangle ABC$ is a right triangle and that AD and thus Ad is the angle bisector for $\angle BAC$. Therefore, we have by our favorite theorem that

$$(AB + AC) : BC = AC : Cd$$

The sum of the cotangent and the cosecant for 2θ is equal to the cotangent of θ . This is the same relationship we used in the first section, with the addition of the similarity between triangles.

We also have that $\triangle ABD$ is a right triangle and by virtue of the angle bisector construction, $\triangle ABD$ is similar to $\triangle ACd$. Therefore:

$$AC : Cd = AD : DB$$

These two lines combine to give the desired result.

□

Note: we did not use the fact from the source that the small triangle $\triangle BDd$ is similar to $\triangle ACd$ (vertical angles plus the two right angles). (To do: go back to Aristotle or

come up with a new proof using the same givens as the bisector theorem plus this additional point, to show the result directly).

In any event, what we will do at each stage is exactly the same as before: (i) add the cotangent and cosecant of 2θ to obtain the cotangent of θ , (ii) then use the Pythagorean theorem to get the cosecant of θ .

We have that the initial cotangent $AC : BC < 1351 : 780$, and $AB : BC$ is the cosecant, whose value is 2 so we multiply $780 \times 2 = 1560$, and then add that to 1351 to get the numerator of the result listed above. This is the cotangent of the half-angle.

$$\circ AB : BD < 3013 \frac{3}{4} : 780$$

From the Pythagorean theorem: $AD^2 + BD^2 = AB^2$ so

$$AB^2 : BD^2 = AD^2 : BD^2 + 1$$

$AD : DB < 2911 : 780$ So we obtain

$$AD^2 = 2911^2 = 8473921$$

$$DB^2 = 780^2 = 608400$$

$$AD^2 + BD^2 = 9082321$$

$$AB = 3013 - 3/4$$

$$AB : BD < 3013 \frac{3}{4} : 780$$

I2

Rather than go through the geometry again, let's just use the trigonometry. First the cotangent: we have $2911 : 780 + 3013 \frac{3}{4} : 780 = 5924 \frac{3}{4} : 780$.

At this point, we reduce the denominator to 240. This amounts to dividing by $3 \frac{1}{4}$. $5924 \frac{3}{4}$ divided by $3 \frac{1}{4}$ is exactly equal to 1823.

$$\circ AE : EB = 1823 : 240, \text{ the cotangent.}$$

For the second part

$$1823^2 = 3323329$$

$$240^2 = 57600$$

$$AE^2 + BE^2 = 3380929$$

$$AE/EB = 3013 - 3/4$$

The square root is $< 1838 \frac{3}{4}$, but the source gives the fraction as a bit larger
 $1838 \frac{9}{11}$.

- $AB : BE = 1838 \frac{9}{11} : 240$, the cosecant.

I3

Now, let AF bisect the angle.

For the first part we have $1838 \frac{9}{11} : 240 + AE : EB = 1823 : 240 = 3661 \frac{9}{11} : 240$.

We reduce the denominator, this time to 66. This amounts to multiplication by $11/40$. So the numerator is multiplied by the same factor giving

- $AF : FB = 1007 : 66$, the cotangent.

$$1007^2 = 1014049$$

$$66^2 = 4356$$

$$AF^2 + BF^2 = 1018405$$

$$AB/FB = 1009-1/6$$

- $AB : FB = 1009 \frac{1}{6} : 66$, the cosecant.

I4

Finally, let AG bisect the angle, and then join BG . First the cotangent

$$(AB + AF) : BF = AG : GB$$

- $AG : GB = 2016 \frac{1}{6} : 66$

$$2016^2 = 4064256$$

$$66^2 = 4356$$

$$AG^2 + BG^2 = 4068612$$

$$AB/GB = 2017-1/12$$

The source gives

- $AB : GB < 2017 \frac{1}{4} : 66$. This is the cosecant.

We're almost done. The side BG is a side of an inscribed regular polygon of 96 sides.
We multiply $66 \times 96 = 6336$ and compute the inverse ratio to $2017 \frac{1}{4}$.

I am not sure how Archimedes came up with it, but it is easy to verify that the ratio which is less than π is greater than:

$$\frac{6336}{2017 \frac{1}{4}} > 3 \frac{10}{71}$$

We combine parts A and B in the famous final statement that

$$3 \frac{10}{71} < \pi < 3 \frac{1}{7}$$

Check our progress

Here are calculations for the lower bound at the beginning and then after each of four steps:

```
>>> 6 * 1/2.0
3.0
>>> 12*780/3013.75
3.1057652426379097
>>> 24*240/1838.82
3.1324436323294287
>>> 48*66/1009.1666
3.1392239893789586
>>> 96*66/2017.25
3.1409096542322468
>>>
>>> 10.0/71
0.14084507042253522
```

Chapter 33

Value of pi

Archimedes and π

Since Archimedes is a strong presence in this book, we will discuss his method for calculating an approximation to the value of π , the ratio of the circumference of a circle to its diameter.

The commonly cited result is

The ratio of the circumference of any circle to its diameter is less than $3\frac{1}{7}$ but greater than $3\frac{10}{71}$.

In decimal that is

$$3.140845\dots < \pi < 3.1428571$$

However, while useful, this misses the main idea: Archimedes described an iterative procedure which can be used to calculate the value of π *to any desired accuracy*.

Although the idea is beautiful, his argument is somewhat unwieldy in detail, so instead we will use modern trigonometry to achieve the same result more economically.

For a discussion of Archimedes actual method (based on a translation by Heath), see this web page

<https://itech.fgcu.edu/faculty/clindsey/mhf4404/archimedes/archimedes.html>

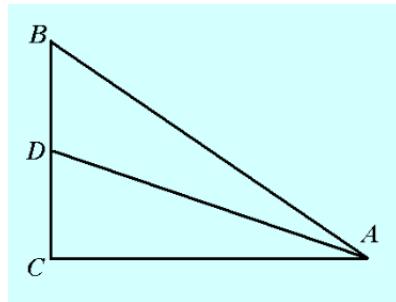
and I have worked out the same proof in detail in the book Best of Calculus.

https://github.com/telliott99/calculus_book

We will also use the trigonometry to find easy formulas for the perimeter and area of inscribed and circumscribed polygons. That part of the argument is partly in this chapter, and the second part follows.

angle bisector

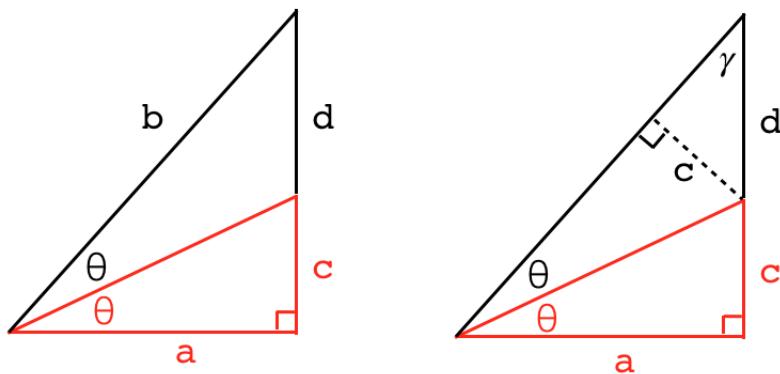
Although we don't follow Archimedes exactly, a key element which he relies upon is the proof that, for an angle bisector in a right triangle, the adjacent sides are in the same proportion as the two segments formed where the bisector meets the other side.



Here:

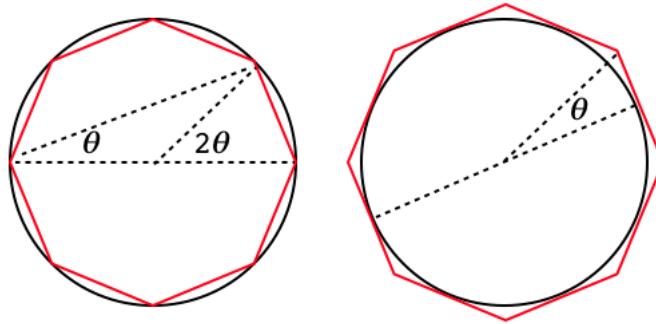
$$\frac{AB}{AC} = \frac{BD}{DC}$$

We showed a proof of this ([here](#)), or you may be able to reconstruct it from the figure:



the method

We approximate the value of π by squeezing it between the perimeter of an inscribed polygon, which is less than the circumference of the circle, and the perimeter of a circumscribed polygon, which is greater than the circumference of the circle.



The circle has *diameter* equal to 1 (rather than radius 1, which is more usual). The circumference of the circle is then equal to π , the value which gets squeezed between the two perimeters.

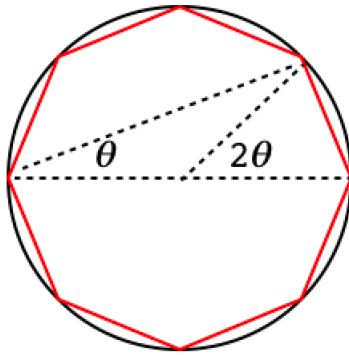
The figure shows a sketch of the polygons when $n = 8$. We will be increasing the number of sides by a factor of 2 at each step, so these are really 2^n -gons with $n = 3$ here.

Finding perimeters in terms of angle θ

For the inscribed circle (left panel), there are 8 sides, so the central angle (marked 2θ) is equal to

$$\frac{2\pi}{8} = \frac{\pi}{4} = 45^\circ$$

and θ is one-half that.

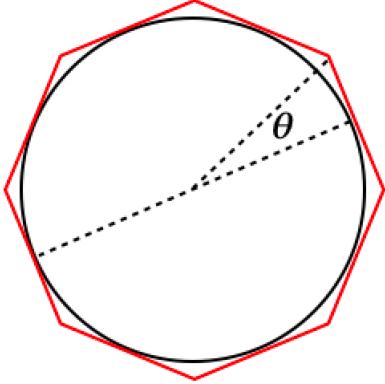


By Thales' theorem, the triangle above containing angle θ , with the diameter as one side, and two other vertices also on the circle, is a right triangle.

The inscribed n -gon side of length S (shown in red) is equal to $\sin \theta$, since the hypotenuse of the triangle is the diameter of the circle, which is equal to 1.

The total perimeter is $8 \cdot S$.

Alternatively, use half the angle at the center of the circle (i.e. θ). Then half the length of the red line $S/2$, divided by the radius ($r = 1/2$) gives $S = \sin \theta$, the same result.



We have the same circle (now showing the outside polygon, circumscribing the circle), it is just rotated slightly. One dashed line extends a bit further to the vertex of the n -gon outside. The angle marked θ is one-half the angle we marked as 2θ previously since now the diameter comes down to the middle of the side.

We compute the whole length of the side T as follows. The half-side is $T/2$ and the hypotenuse of the triangle is one-half the unit diameter, which is $1/2$, so $T = \tan \theta$.

The total perimeter is $8 \cdot T$.

All of this gives us two simple equations for the two perimeters. At each stage there are 2^n sides, the length of each short side S on the inside equals $\sin \theta$ and the length of each short side on the outside T is equal to $\tan \theta$, where $\theta = 2\pi/2^n$.

The total length of the inside perimeter is

$$p = nS = n \sin \theta$$

and that of the outside is

$$P = nT = n \tan \theta$$

When we go from θ to $\theta' = \theta/2$ and $n' = 2n$, we must compute the new values S' and T' from S and T using the half-angle formulas, and then also multiply by 2 to take account of the change from n to $2n$ for the total circumference.

The base case

If we go back to the square ($n = 2$, $2^n = 4$), then the angle θ is $\pi/4$.

The tangent is $T = \tan \pi/4 = 1$ and the sine is $S = \sin \pi/4 = 1/\sqrt{2}$.

Our formulas say that on the inside, the perimeter is $4S = 4/\sqrt{2} = 2\sqrt{2}$ and on the outside, the perimeter is $4T = 4$.

We can just check that from simple geometry. Calculate that the circumscribing square has a side length which is twice the radius of the circle, that is, $s = 1$ for our circle with unit diameter, so its perimeter is 4, which is correct.

Similarly, an inscribed square can be decomposed into four isosceles right triangles with sides of length $1/2$ and hypotenuse $1/\sqrt{2}$, so the total perimeter is $4/\sqrt{2}$, which also checks.

Now, what we will do is to increase n in steps of 1, that increases 2^n by a factor of $2^1 = 2$ each time. Doubling n halves the angle. So all we need is a way to compute trigonometric functions of $\theta/2$, knowing the values for θ , and then we can calculate what happens to the perimeter.

π is simply that value, since the diameter is 1.

Half angle formulas

Quick review ([here](#)).

Let unprimed values refer to the whole angle, while primed ones refer to the half-angle. Then:

$$\begin{aligned} S &= 2S'C' \\ C + 1 &= 2[C']^2 \\ \frac{1}{T'} &= \frac{1}{T} + \frac{1}{S} \end{aligned}$$

and

$$T' = \frac{1}{1/S + 1/T} = \frac{S}{1+C}$$

So, given S, C and T , we can calculate T' and C' , and then finally S' . To get the perimeters, remember that factor of two from doubling n , the number of sides.

Calculation

Let's run a simulation to see what kind of numbers we get. Start with the square ($n = 2, 2^n = 4$)

Previously we found that $S = 1/\sqrt{2}$ and $T = 1$ so

$$\begin{aligned} p &= 2^n S = \frac{4}{\sqrt{2}} = 2.8284 \\ P &= 2^n T = 4 \end{aligned}$$

Let's try a script to calculate this to larger n .

<https://gist.github.com/telliott99/19f521c807210171a4847b319104b3df>

Output:

```
> python pi.py
2 2.8284271247 4.0000000000
3 3.0614674589 3.3137084990
4 3.1214451523 3.1825978781
```

```
5 3.1365484905 3.1517249074
6 3.1403311570 3.1441183852
7 3.1412772509 3.1422236299
8 3.1415138011 3.1417503692
9 3.1415729404 3.1416320807
10 3.1415877253 3.1416025103
11 3.1415914215 3.1415951177
12 3.1415923456 3.1415932696
13 3.1415925766 3.1415928076
14 3.1415926343 3.1415926921
15 3.1415926488 3.1415926632
16 3.1415926524 3.1415926560
17 3.1415926533 3.1415926542
18 3.1415926535 3.1415926537
19 3.1415926536 3.1415926536
```

>

That looks pretty good to me, although it's a bit slow to converge.

This is really quite amazing. Archimedes has not only calculated π to 3 significant figures. More important, he has provided us with an iterative procedure that can be used to calculate the value to *any precision we desire*. As an engineer, Archimedes knew that 3.1416 is precise enough, so he stopped.

After all, no one wants to be William Shanks, or one of these guys:

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

Quote:

[He] calculated pi to [n] digits, but *not all were correct*.

alternative approach to the perimeter

This web page originally got me started with this derivation

<http://personal.bgsu.edu/~carother/pi/Pi3d.html>

(Unfortunately, the link is dead now, probably because the University took Dr. Carother's pages down). It has been preserved by the wayback machine:

<https://web.archive.org/web/20171024182015/http://personal.bgsu.edu/~carother/pi/Pi3d.html>

On that page, there was given a different pair of formulas, namely, for an inside perimeter p and an outside perimeter P

$$P' = \frac{2pP}{p + P}$$

$$p' = \sqrt{pP'}$$

The first equation can be rearranged to give

$$\frac{1}{P'} = \frac{1}{2} \left[\frac{1}{P} + \frac{1}{p} \right]$$

which is the definition of the harmonic mean of p and P , while the second equation is the geometric mean.

Since in our derivation p and P are the same multiple of S and T , the same relationships should hold for the sine and tangent, but we must remember the extra factor of 2.

From the double angle formula, we said that

$$\frac{1}{T'} = \frac{1}{S} + \frac{1}{T}$$

Recall that S is the same as p , within a factor of n , and that T is the same as P , within the same factor.

$$p = nS$$

$$P = nT$$

while

$$P' = 2nT'$$

So we can rewrite the above equation as

$$\frac{2}{P'} = \frac{1}{p} + \frac{1}{P}$$

$$\frac{1}{P'} = \frac{1}{2} \left(\frac{1}{p} + \frac{1}{P} \right)$$

The harmonic mean. Or

$$P' = 2 \frac{p+P}{pP}$$

This is what was given.

□

For the second one

$$S' = \frac{S}{2C'} = \frac{S}{2} \frac{T'}{S'}$$

Then

$$2[S']^2 = S \cdot T'$$

$$4[S']^2 = S \cdot 2T'$$

$$[2nS']^2 = nS \cdot 2nT'$$

Changing variables, $p' = 2nS'$

$$[p']^2 = pP'$$

Finally

$$p' = \sqrt{pP'}$$

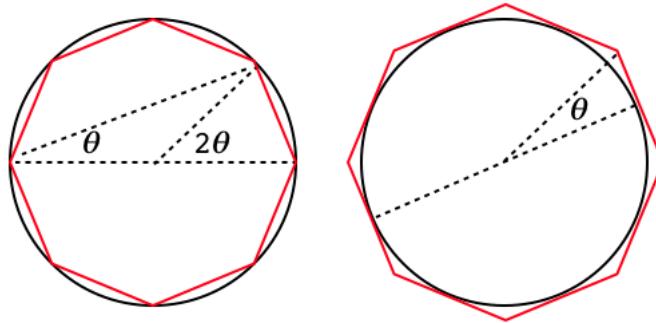
which matches what was given.

□

area

There is yet another way to apply the method, and that is to calculate the *areas* of inscribed and circumscribed polygons. Let's go through this briefly here and look more carefully at the geometry in the next chapter.

For this approach we use a **unit circle** (radius 1) rather than a diameter of 1, as we did above. As before, we define θ to be the central angle of the half-sector (i.e. $\theta = 2\pi/2n$).



Rather than draw an entirely new figure, just imagine in the left panel that we draw the angle bisector of angle 2θ .

Each half triangle has base $\cos \theta$ and height $\sin \theta$, but there are two of them, so the total area of the whole triangle is just $\sin \theta \cos \theta$ and the total area of the inner polygon is

$$a = n \sin \theta \cos \theta = nSC$$

in the notation of this chapter.

As before, to progress to a' we have a factor of 2 as well as the new values S' and C' :

$$a' = 2nS'C'$$

For the circumscribed or outer polygon, we just have what we had before, that the side length of the triangle in the right panel is $\tan \theta$ so the total area is

$$A = nT$$

Bring in the half-angle formulas as follows:

$$a' = 2nS'C' = 2n \cdot \frac{S}{2C'} \cdot C' = nS$$

That is slick, but we need an expression for nS :

$$aA = nSC \cdot n \frac{S}{C} = [nS]^2$$

$$aA = [a']^2$$

$$a' = \sqrt{aA}$$

This is like, and yet subtly different than what we had when calculating the perimeter.

Since

$$A = nT$$

and

$$\begin{aligned} A' &= 2nT' \\ &= 2n \frac{ST}{S+T} = 2 \frac{nS \cdot nT}{nS + nT} \\ A' &= 2 \frac{a'A}{a'+A} \end{aligned}$$

Compare

$$\begin{aligned} a' &= \sqrt{aA} & A' &= 2 \frac{a'A}{a'+A} \\ p' &= \sqrt{pP'} & P' &= 2 \frac{pP}{p+P} \end{aligned}$$

We have primed values in corresponding positions.

However, it turns out that when you take account of the differing size of the circle for perimeter and area methods, and thus the initial values of p , P , a and A , the different order of operations results in precisely the same calculation.

historical note

The area-based formulas given above are due to James Gregory.

<https://divisbyzero.com/2018/09/28/proof-without-word-gregorys-theorem/>

As an aside, the Fundamental Theorem of Calculus (FTC) is usually thought about (taught and learned) using the language of functions, and ascribed mainly to Leibnitz, with some credit to the two Isaacs, Newton and his university lecturer, Barrow.

<https://arxiv.org/abs/1111.6145>

Amazingly enough, Gregory published a geometric (Euclidean) proof of the FTC in 1668! That predates Liebnitz (1693) by more than 25 years. This is motivation to give considerable credit to individuals other than Newton and Liebnitz (e.g. Fermat, Pascal, Wallis, Gregory, etc.) in the invention of the calculus.

test

I wrote a simple test of the area formulas using Python.

The script is here:

<https://gist.github.com/telliott99/5269b48672cdaeca95c9c9d163321d>

It gives this output:

```
> python script.py
 4 2.0000000000 4.0000000000
 8 2.8284271247 3.3137084990
16 3.0614674589 3.1825978781
32 3.1214451523 3.1517249074
64 3.1365484905 3.1441183852
128 3.1403311570 3.1422236299
256 3.1412772509 3.1417503692
512 3.1415138011 3.1416320807
1024 3.1415729404 3.1416025103
2048 3.1415877253 3.1415951177
4096 3.1415914215 3.1415932696
8192 3.1415923456 3.1415928076
16384 3.1415925766 3.1415926921
32768 3.1415926343 3.1415926632
65536 3.1415926488 3.1415926560
>
```

The digits of the output appear to be identical or nearly so. The only difference is that in this script I computed 2^n to give the number of sides. In the previous chapter, we just print n .

details

That's very curious. The first four lines of output from the perimeter version:

```
4 2.8284271247 4.0000000000
8 3.0614674589 3.3137084990
16 3.1214451523 3.1825978781
32 3.1365484905 3.1517249074
```

and the first five from the area version:

4	2.0000000000	4.0000000000
8	2.8284271247	3.3137084990
16	3.0614674589	3.1825978781
32	3.1214451523	3.1517249074
64	3.1365484905	3.1441183852

It's pretty clear that we are doing the same calculation. It's just that the first column is shifted up by one row.

To confirm that, the perimeter calculation is:

initialization:

$$p = 2\sqrt{2} \quad P = 4$$

recurrence:

$$P' = \frac{2pP}{p+P} \quad p' = \sqrt{pP'}$$

The area version is:

initialization:

$$a = 2 \quad A = 4$$

recurrence:

$$a' = \sqrt{aA} \quad A' = \frac{2a'A}{a'+A}$$

They give identical results: $A = P$, at each round, but a matches p' , or to put it the other way around, p' is retarded by one cycle compared to a' .

Let's try one round of calculation by hand. I found it much easier to start with a square.

$$\begin{aligned} p &= 2\sqrt{2} \quad P = 4 \\ P' &= \frac{2pP}{p+P} = \frac{2 \cdot 2\sqrt{2} \cdot 4}{2\sqrt{2} + 4} = \frac{2 \cdot 2\sqrt{2} \cdot 4}{2\sqrt{2}(1 + \sqrt{2})} = \frac{8}{1 + \sqrt{2}} = 3.31371 \\ p' &= \sqrt{pP'} = \sqrt{2\sqrt{2} \cdot \frac{8}{1 + \sqrt{2}}} = 4\sqrt{\frac{1}{1 + 1/\sqrt{2}}} = 3.06147 \end{aligned}$$

The area calculation:

$$a' = \sqrt{aA} = \sqrt{2 \cdot 4} = \sqrt{8} = 2.828427$$

$$A' = \frac{2a'A}{a' + A} = \frac{2 \cdot \sqrt{8} \cdot 4}{\sqrt{8} + 4} = \frac{8}{1 + \sqrt{2}}$$

A' is the same as P' .

The next round for a' is

$$a' = \sqrt{aA} = \sqrt{\sqrt{8} \cdot \frac{8}{1 + \sqrt{2}}} = 4\sqrt{\frac{1}{1 + 1/\sqrt{2}}}$$

Chapter 34

Value of pi revisited

As discussed in a previous [chapter](#), Archimedes used paired inscribed and circumscribed polygons to develop an iterative procedure that can be used to calculate the value of π *to any desired accuracy*.

Although the method is beautiful, his argument is unwieldy in detail, so we used modern trigonometry to achieve the same result more economically.

There are, in addition, two other sets of formulas that also reach this end, one based on perimeters, and the other on areas. These formulas are intriguing because they are simple, and it is not surprising that they are connected.

For example, consider a circle of unit *diameter*, so that π is equal to the perimeter. If p and P are the inside and outside perimeters for polygons whose sectors have central angle θ , and the same symbols are used with primes for angle $\theta/2$, then:

$$P' = 2 \frac{pP}{p + P}$$
$$p' = \sqrt{pP'}$$

The corresponding formulas for inside (a) and outside (A) areas are (for a circle of unit radius)

$$A' = 2 \frac{a'A}{a' + A}$$
$$a' = \sqrt{aA}$$

Notice that these two similar sets of formulas are subtly different. For example, to go from p and P to the primed version, we start with the first formula, while for area we must start with the square root. Part of our purpose in this chapter is to show that this works.

inspiration

It's striking that the formulas for the inside and outside perimeters are so simple, namely $n \sin \theta$ and $n \tan \theta$. The rest just follows from the half-angle formulas.

The web page which originally got me started with the harmonic and geometric mean formulas has been preserved by the wayback machine:

<https://web.archive.org/web/20171024182015/http://personal.bgsu.edu/~carother/pi/Pi3d.html>

On the very same day that I was revising the previous chapter to better integrate these two approaches, I came across another page which gives a "proof without words" of Gregory's Theorem (that is our subject).

<https://divisbyzero.com/2018/09/28/proof-without-word-gregorys-theorem/>

It gives these two formulas:

$$\begin{aligned} I_{2n} &= \sqrt{I_n C_n} \\ C_{2n} &= \frac{2}{1/I_{2n} + 1/C_n} \end{aligned}$$

I found this notation a bit awkward, so I substituted the versions given above:

$$a' = \sqrt{aA}$$

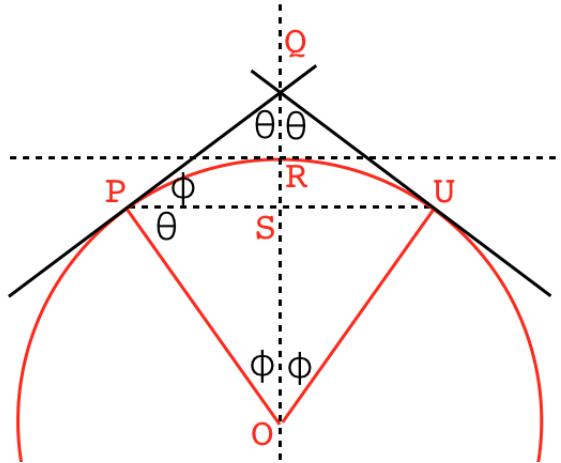
$$A' = 2 \frac{a' A}{a' + A}$$

Here, we mainly follow the development from that page and its "proof without words".

One difference is that we will start with the geometry and work backward to the formulas. Another is that we will use quite a few words. Let's deal with the perimeter first and then do the area.

chord of a circle

Before we start the main part, let's just establish some facts about a chord of a circle.



Draw three radii of the circle such that the third bisects the first two, with angles equal to ϕ at the center, point O .

We will show that the chord is parallel to the tangent at R , and perpendicular to OSQ .

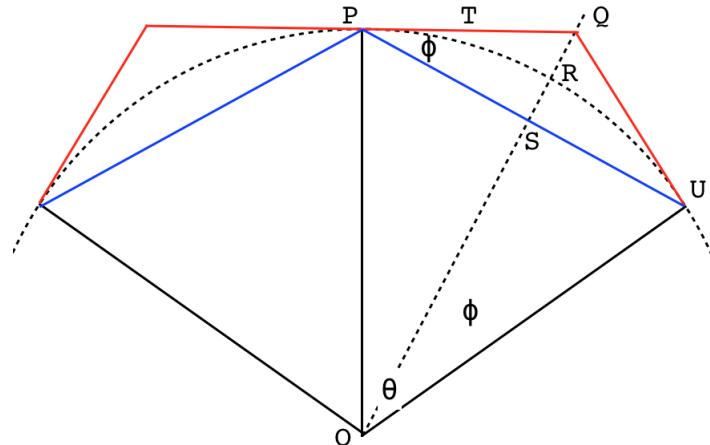
First, tangents are always perpendicular to the radius at the point where they meet the circle.

$\triangle OPS \cong \triangle OSU$ by SAS. Therefore, all four angles at S are right angles. Therefore the chord PU is perpendicular to $OSRQ$. Therefore the chord is parallel to the tangent at R .

The other angle labels are justified as complementary angles in a right triangle.

So any chord which is bisected in this way, is parallel to the tangent above it. And the bisection of the chord follows from the bisection of the base angle at O .

basic setup



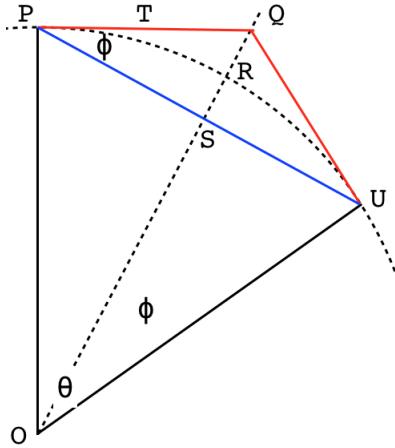
Draw a circle centered at O (only a part of the circle is shown).

Divide the whole 2π radians into $2n$ parts such that ϕ is equal to one of these parts and θ is equal to two of them.

Draw OP , OR and OU as radii, and extend OR to OQ .

Draw the sides of a regular polygon with n sides inscribing the circle and touching its perimeter at P and U . Similarly draw the sides of an n -gon circumscribing the circle and touching its perimeter at the same points P and U .

Two red lines comprise this sector's external perimeter P , while a single blue line is the inscribed perimeter p . The lines of the external perimeter are both tangent to the circle at P and U , and the whole figure is symmetric in each sector, with one blue and two red lines.



By our preliminary results, $\angle PSR$ is a right angle and $\angle SPQ$ is equal to ϕ .

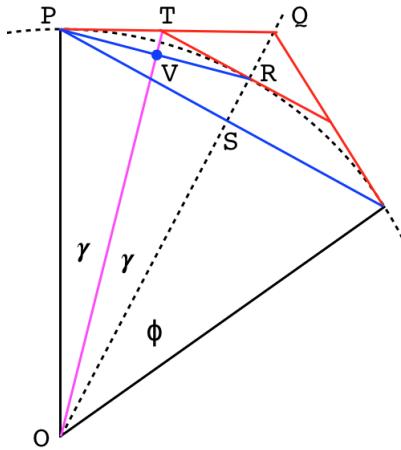
And as we said, the perpendicular bisector to a chord is perpendicular to the tangent at the point where the chord intersects the circle (R).

bisecting ϕ

Next, draw the perimeters p' and P' for the polygon with $2n$ sides and sector angle $\phi = \theta/2$.

It is convenient to rotate the internal perimeter by $\theta/2$ with respect to the external one, a bit to the left when we draw p' and a bit to the right for P' . Both p' and P' touch the circle at R .

A central relationship we use below is that $\triangle PRT$ is isosceles.



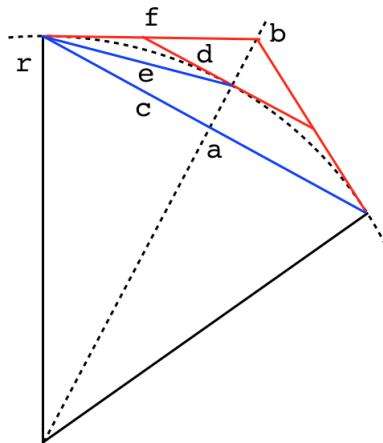
Proof.

$\angle VPT \cong \angle VRT$ and the angles at V are right angles, by the preliminary result. Therefore $\angle TPV = \angle TRV$ so $\triangle PRT$ is isosceles. By complementary angles, the base angle has measure γ .

□

It looks as if the segment of the vertical that extends beyond the radius might be equal to that part below down to what looks like the "strut" of a kite. However, this is not true. We will show what this ratio is equal to in just a bit.

Rather than use the vertices as points of reference, let us now label the line segments.



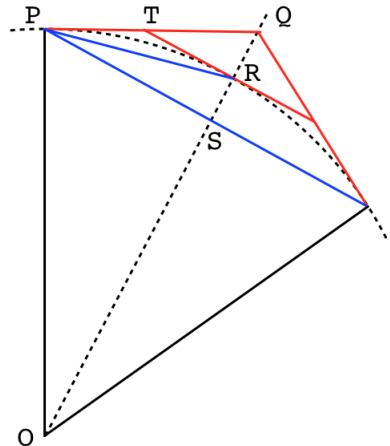
Just to be clear: a is the part of the radius extended to point S in the previous drawing, the intersection of the dashed black and solid blue lines, while b extends all the way to Q , at the vertex of the circumscribing polygon.

c and d are the lengths of the indicated lines *in the half-sector*, not all the way across, and f is the entire length of PQ .

We're ready to proceed.

perimeters

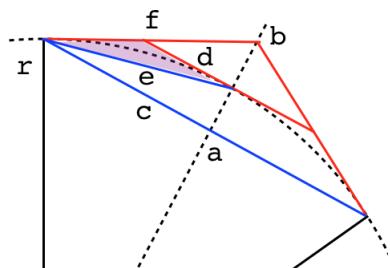
As we said, the key observation is that $\triangle PRT$ is isosceles.



Because of that, and since $\angle SPR = \angle PRT$ by the alternate interior angles theorem, $\angle SPR = \angle TPR$.

Therefore the cosines are also equal, namely:

$$\frac{c}{e} = \frac{e/2}{d}$$



(To see the midpoint of e , drop an altitude in the isosceles triangle, shown in purple).

Therefore:

$$2dc = e^2$$

Now, c is the entirety of p in this half-sector. But d is only one-half of P' .

Hence $2d \cdot c$ is equal to pP' , and since $e = p'$, we have that

$$pP' = [p']^2$$

which was our second rule for the perimeters.

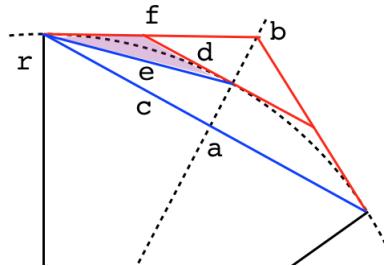
The first rule was

$$P' = 2 \frac{pP}{p + P}$$

In geometric terms, we must show that

$$\begin{aligned} 2d &= 2 \frac{cf}{c + f} \\ cd + df &= cf \end{aligned}$$

Taking another look at the diagram:



The small triangle with base d ($\triangle QRT$ above) has slanted side $f - d$ (subtracting d because, again, $\triangle PRT$ is isosceles). By similar triangles, we have

$$\begin{aligned} \frac{d}{f - d} &= \frac{c}{f} \\ df &= cf - cd \\ cd + df &= cf \end{aligned}$$

Which is what we needed to prove.

□

areas

The area formulas for inside (a) and outside (A) polygons are those for a circle of unit radius (so that π is the area):

$$A' = 2 \frac{a' A}{a' + A}$$

$$a' = \sqrt{a A}$$

This is what we will prove.

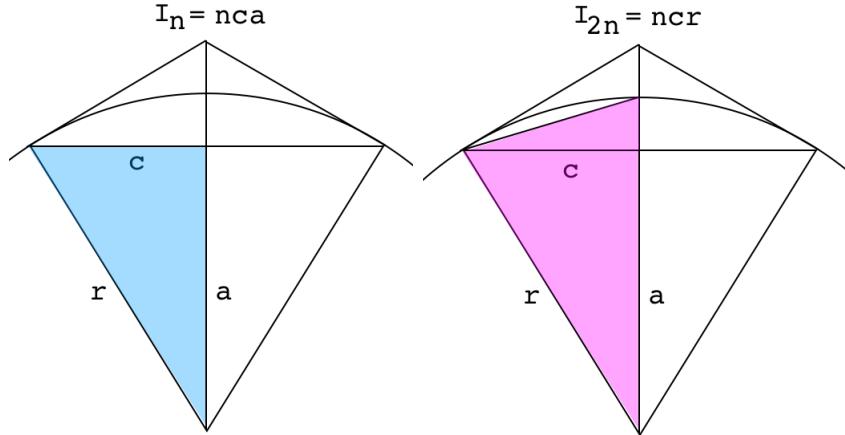
However, having reached this point, we need another symbol for area, because a is currently the line segment corresponding to p/n . Let's use I and C for the inside and outside areas, to match the source.

We will also adopt their n and $2n$ notation, It's a bit clumsy but that will make it easier to match things up. Substituting in the above equations:

$$C_{2n} = 2 \cdot \frac{I_{2n} C_n}{I_{2n} + C_n}$$

$$I_{2n} = \sqrt{I_n C_n}$$

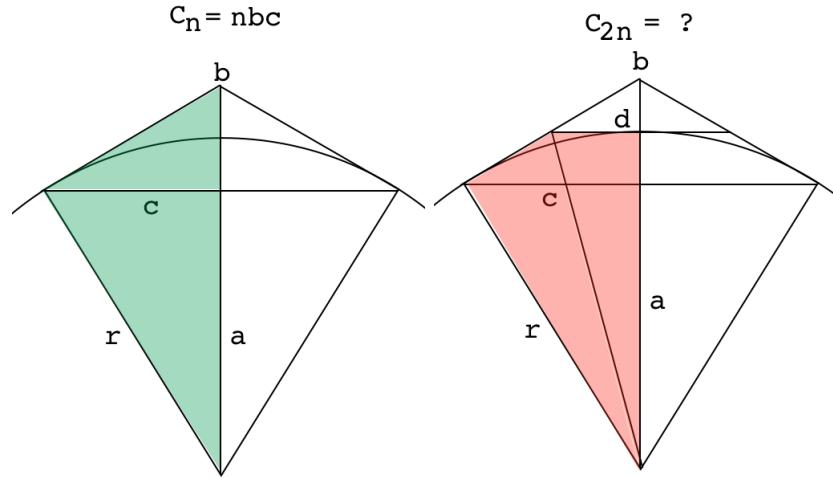
The first two areas are I_n and I_{2n}



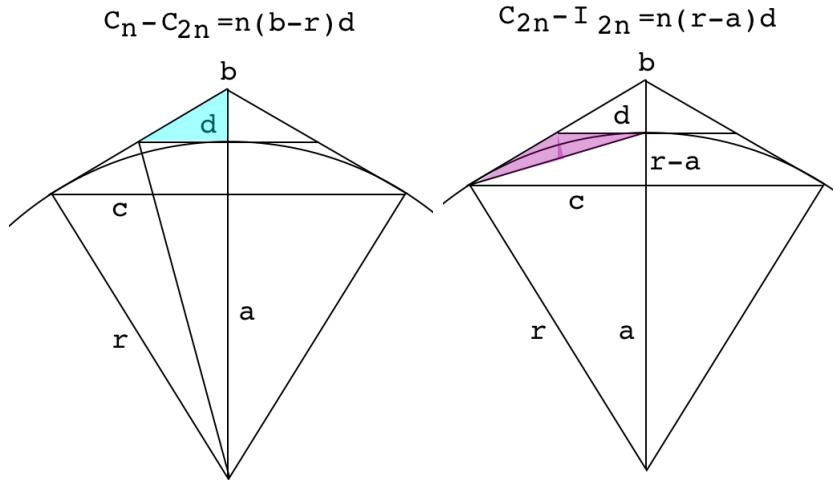
We compute these areas for the whole sector of angle θ , so there are two congruent triangles with base a (or base r) and height c , which makes the factors of one-half go away.

Multiply by n if you like to get the entire polygon, but every expression will have a factor of n , and we'll be looking at ratios, so we can just not worry about it.

The third easy one is C_n :



We write the last one (C_{2n}) as two different differences.



Let's gather all these expressions in one place, forming ratios:

$$\frac{I_{2n}}{I_n} = \frac{ncr}{nca} = \frac{r}{a}$$

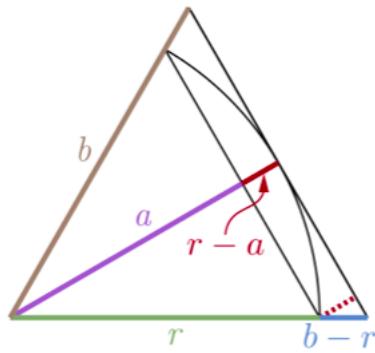
$$\frac{C_n}{I_{2n}} = \frac{ncb}{ncr} = \frac{b}{r}$$

$$\frac{C_n - C_{2n}}{C_{2n} - I_{2n}} = \frac{n(b-r)d}{n(r-a)d} = \frac{b-r}{r-a}$$

We will prove that these three ratios are all equal to each other.

We have used the geometry to prove what the source calls their Lemmas, and those can be used in turn to prove the original Gregory formulas.

But the proof is easy:



It's just a matter of similar triangles:

$$\frac{r}{a} = \frac{b}{r} = \frac{b-r}{r-a}$$

That's the "without words" part.

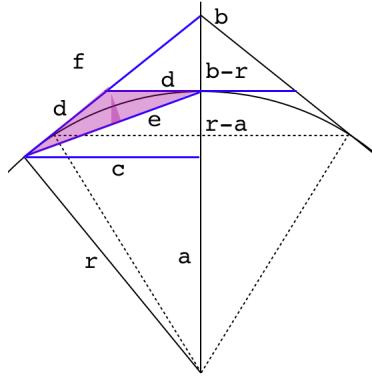
For that very last part, you can work out the dimensions of the tiny similar triangle, or you can say:

$$\begin{aligned}\frac{r}{a} &= \frac{b}{r} \\ \frac{r}{a} - \frac{a}{a} &= \frac{b}{r} - \frac{r}{r} \\ \frac{r-a}{a} &= \frac{b-r}{r}\end{aligned}$$

which is easily rearranged to give the desired result.

□

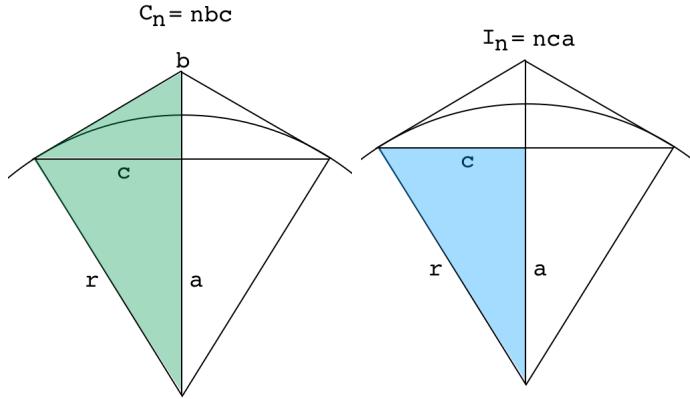
This can also be proved using the **angle bisector theorem**.



The side labeled e bisects the angle formed by the two sides labeled c and f . Therefore

$$\frac{b-r}{f} = \frac{r-a}{c} \Rightarrow \frac{b-r}{r-a} = \frac{f}{c}$$

But f and c are two sides of a triangle which is similar to the colored portions below:



Therefore

$$\frac{b}{r} = \frac{r}{a} = \frac{f}{c} = \frac{b-r}{r-a}$$

As we said.

algebra

Moving on to the geometric mean formula is not hard. From above we have that

$$\frac{I_{2n}}{I_n} = \frac{C_n}{I_{2n}}$$

$$[I_{2n}]^2 = I_n C_n$$

Translated back into the A, a area notation

$$a' = \sqrt{aA}$$

This is just what we wanted to show.

For the other formula, what we have is:

$$\frac{C_n - C_{2n}}{C_{2n} - I_{2n}} = \frac{C_n}{I_{2n}}$$

$$\begin{aligned} I_{2n}(C_n - C_{2n}) &= C_n(C_{2n} - I_{2n}) \\ 2I_{2n}C_n &= C_nC_{2n} + I_{2n}C_{2n} \\ &= C_{2n}(C_n + I_{2n}) \end{aligned}$$

So

$$\begin{aligned} C_{2n} &= 2 \cdot \frac{I_{2n}C_n}{C_n + I_{2n}} \\ C_{2n} &= 2 \cdot \frac{1}{1/I_{2n} + 1/C_n} \end{aligned}$$

And we're done. In our preferred notation

$$A' = 2 \cdot \frac{1}{1/a' + 1/A}$$

historical note

The area-based formulas given above are due to James Gregory.

<https://divisbyzero.com/2018/09/28/proof-without-word-gregorys-theorem/>

As an aside, the Fundamental Theorem of Calculus (FTC) is usually thought about (taught and learned) using the language of functions, and ascribed mainly to Leibnitz, with some credit to the two Isaacs, Newton and his university lecturer, Barrow.

<https://arxiv.org/abs/1111.6145>

Amazingly enough, Gregory published a geometric (Euclidean) proof of the FTC in 1668! That predates Liebnitz (1693) by more than 25 years. This is motivation to give considerable credit to individuals other than Newton and Liebnitz (e.g. Fermat, Pascal, Wallis, Gregory, etc.) in the invention of the calculus.

test

I wrote a simple test of the area formulas using Python.

The script is here:

<https://gist.github.com/telliott99/5269b48672cdaeca95c9c9d163321d>

It gives this output:

```
> python script.py
 4 2.0000000000 4.0000000000
 8 2.8284271247 3.3137084990
16 3.0614674589 3.1825978781
32 3.1214451523 3.1517249074
64 3.1365484905 3.1441183852
128 3.1403311570 3.1422236299
256 3.1412772509 3.1417503692
512 3.1415138011 3.1416320807
1024 3.1415729404 3.1416025103
2048 3.1415877253 3.1415951177
4096 3.1415914215 3.1415932696
8192 3.1415923456 3.1415928076
16384 3.1415925766 3.1415926921
32768 3.1415926343 3.1415926632
65536 3.1415926488 3.1415926560
>
```

The digits of the output appear to be identical or nearly so. The only difference is that in this script I computed 2^n to give the number of sides. In the previous chapter, we just print n .

details

That's very curious. The first four lines of output from the perimeter version:

```
 2 2.8284271247 4.0000000000
 3 3.0614674589 3.3137084990
 4 3.1214451523 3.1825978781
 5 3.1365484905 3.1517249074
```

and the first five from the area version:

4	2.0000000000	4.0000000000
8	2.8284271247	3.3137084990
16	3.0614674589	3.1825978781
32	3.1214451523	3.1517249074
64	3.1365484905	3.1441183852

It's pretty clear that we are doing the same calculation. It's just that the first column is shifted up by one row.

To confirm that, the perimeter calculation is:

initialization:

$$p = 2\sqrt{2} \quad P = 4$$

recurrence:

$$P' = \frac{2pP}{p+P} \quad p' = \sqrt{pP'}$$

The area version is:

initialization:

$$a = 2 \quad A = 4$$

recurrence:

$$a' = \sqrt{aA} \quad A' = \frac{2a'A}{a'+A}$$

They give identical results: $A = P$, at each round, but a matches p' , or to put it the other way around, p' is retarded by one cycle compared to a' .

Let's try one round of calculation by hand:

$$\begin{aligned} p &= 2\sqrt{2} \quad P = 4 \\ P' &= \frac{2pP}{p+P} = \frac{2 \cdot 2\sqrt{2} \cdot 4}{2\sqrt{2} + 4} = \frac{2 \cdot 2\sqrt{2} \cdot 4}{2\sqrt{2}(1 + \sqrt{2})} = \frac{8}{1 + \sqrt{2}} = 3.31371 \\ p' &= \sqrt{pP'} = \sqrt{2\sqrt{2} \cdot \frac{8}{1 + \sqrt{2}}} = 4\sqrt{\frac{1}{1 + 1/\sqrt{2}}} = 3.06147 \end{aligned}$$

The area calculation:

$$a' = \sqrt{aA} = \sqrt{2 \cdot 4} = \sqrt{8} = 2.828427$$

$$A' = \frac{2a'A}{a' + A} = \frac{2 \cdot \sqrt{8} \cdot 4}{\sqrt{8} + 4} = \frac{8}{1 + \sqrt{2}}$$

A' is the same as P' .

The next round for a' is

$$a' = \sqrt{aA} = \sqrt{\sqrt{8} \cdot \frac{8}{1 + \sqrt{2}}} = 4\sqrt{\frac{1}{1 + 1/\sqrt{2}}}$$

I don't have any words of wisdom to explain why this all dovetails so neatly, but there must be one. When things fit together like this it is never an accident.

Chapter 35

Square root of 3

Archimedes uses two approximations for $\sqrt{3}$: $265/153$ and $1350/780$. (Recall that $\sqrt{3}$ is irrational so it doesn't have an exact decimal representation).

I wrote a script to search for these and other close approximations.

```
> python approx_sqrt3.py
    0      0      0      1      1
    1      1      2      2      1
    3      5      2      6      9
    4      6     12      7      1
   11     19      2     20     37
   15     25     50     26      1
   41     71      2     72    141
   56     96    192     97      1
  153    265      2    266    529
  209    361    722    362      1
  571    989      2    990   1977
  780   1350   2700   1351      1
 2131   3691      2   3692   7381
 2911   5041  10082   5042      1
 7953  13775      2  13776  27549
 10864 18816  37632  18817      1
>
```

The first column has the denominator i (we search from 1 to 50000).

The algorithm is really brute force. For each possible i , we search all the integers larger than i until we find one j such that $3 \cdot i^2 < j^2$. In other words, j is the smallest integer such that j/i is larger than $\sqrt{3}$.

Having the closest j (and $j - 1$) for each i , we test whether

$$j^2 - 3 \cdot i^2 < 5$$

$$3 \cdot i^2 - (j - 1)^2 < 5$$

If either is true, we print all the values e.g.

153	265	2	266	529
-----	-----	---	-----	-----

In the third column we find repeatedly, 2.

What this means is that the square of the value in column 2, plus 2, is exactly three times the square of the value in column 1. For example:

$$153^2 = 23409$$

$$265^2 = 70225$$

$$3 \times 23409 = 70227$$

Since $265^2/(153^2 + 2) = 3$, $265/153$ is just barely less than $\sqrt{3}$.

The error is $2/23409 \approx 8 \times 10^{-5}$.

In column 5 we see the number 1 repeated.

This is the difference between 3 times the square of the value in column 4 and the square of the value in column 1. For example:

780	1350	2700	1351	1
-----	------	------	------	---

$$780^2 = 608400$$

$$1351^2 = 1825201$$

$$3 \times 608400 = 1825200$$

So $1351/780$ is just barely greater than $\sqrt{3}$, the error is $1/608400 \approx 1.6 \times 10^{-6}$.

Continued fractions

There is another way to find such numbers. A continued fraction is an expression like:

$$1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \dots}}}$$

This particular continued fraction is equal to the famous number ϕ .

$$\phi = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \dots}}}$$

But notice, the second term on the right-hand side is $1/\phi$ so we can write

$$\begin{aligned}\phi &= 1 + \frac{1}{\phi} \\ \phi^2 &= \phi + 1\end{aligned}$$

For more on ϕ see [here](#).

Square roots can be represented as continued fractions. We look first at the slightly easier case of $\sqrt{2}$, before tackling $\sqrt{3}$.

$$(\sqrt{2} - 1)(\sqrt{2} + 1) = 1$$

Rearrange to get a substitution we will use again

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

At the same time, add one and subtract one on the bottom right:

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

substitute

$$= \frac{1}{2 + \frac{1}{\sqrt{2} + 1}}$$

Add one and subtract one again and then substitute again

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

Clearly, this goes on forever.

$$\begin{aligned}\sqrt{2} &= 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \dots}}} \\ &= 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \dots}}}\end{aligned}$$

The numerators are all 1, so this is a *simple* continued fraction for $\sqrt{2}$.

The continued fraction representation of $\sqrt{2}$ is $[1 : 2]$, meaning that there is an initial 1 followed by repeated 2's.

This fraction goes on forever (since $\sqrt{2}$ is irrational). To turn this into an approximate decimal representation of $\sqrt{2}$, ignore the Then the last fraction is $5/2$. Invert and add, repeatedly:

$$\begin{aligned}2 + 1/2 &= 5/2 \\ 2 + 2/5 &= 12/5 \\ 2 + 5/12 &= 29/12 \\ 2 + 12/29 &= 71/29 \\ 2 + 29/71 &= 171/71 \\ 2 + 71/171 &= 413/171\end{aligned}$$

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 easily get more.

square root of 3

The continued fraction representation of $\sqrt{3}$ is $[1, 1, 2, 1, 2, \dots]$, which can be shortened to $[1 : (1, 2)]$.

Here is a derivation:

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

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

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

both of which we will use again. However, going further, add and subtract on the bottom right

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

Divide top and bottom by 2

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

and substitute giving

$$= \frac{1}{1 + \frac{1}{\sqrt{3}+1}}$$

That's the end of step 1.

Now, for the second step, we focus on that last fraction

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

Then for step three, we focus again on the last fraction, which is what we worked with in the first part.

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

So now both terms repeat:

$$\begin{aligned} \sqrt{3} - 1 &= \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \frac{1}{2 + \dots}}}} \end{aligned}$$

which is $[1 : (1, 2)]$, as we said.

We can get approximations for $\sqrt{3}$ similar to what we did for $\sqrt{2}$. Unlike previously, here there are two possibilities. We start with either one of

$$1 + \frac{1}{2 + \dots}$$

$$2 + \frac{1}{1 + \dots}$$

and proceed by ignoring the dots.

The first gives

$$\begin{aligned} 1 + 1/2 &= 3/2 \\ 2 + 2/3 &= 8/3 \\ 1 + 3/8 &= 11/8 \\ 2 + 8/11 &= 30/11 \\ 1 + 11/30 &= 41/30 \\ 2 + 30/41 &= 112/41 \\ 1 + 41/112 &= 153/112 \end{aligned}$$

$$1 + 112/153 = 265/153 = 1.732026$$

The actual value is $\sqrt{3} = 1.732051$, to six places. We have four.

The second gives

$$\begin{aligned} 2 + 1 &= 3 \\ 1 + 1/3 &= 4/3 \\ 2 + 3/4 &= 11/4 \\ 1 + 4/11 &= 15/11 \\ 2 + 11/15 &= 41/15 \\ 1 + 15/41 &= 56/41 \\ 2 + 41/56 &= 153/56 \\ 1 + 56/153 &= 209/153 \\ 2 + 153/209 &= 571/209 \\ 1 + 209/571 &= 780/571 \end{aligned}$$

$$1 + 571/780 = 1351/780 = 1.732051$$

The actual value is $\sqrt{3} = 1.732051$, to six places. We have all six.

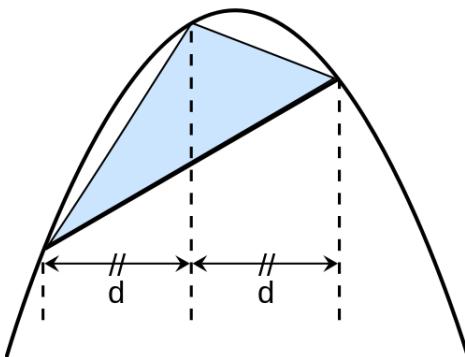
It is believed that this is how Archimedes came up with those approximations. (He doesn't say).

Chapter 36

Archimedes and quadrature

Let's talk about Archimedes, and parabolas.

Here is a figure from wikipedia, showing a parabola and a chord of the parabola, which might be drawn between any two points. A triangle is constructed from the chord in the following way: the point dividing the horizontal distance in half is found and that is used for the x-value of the third point.



The Greek genius Archimedes showed that the total area underneath the curve, between the two outside vertices of the triangle, is $4/3$ times the area of the triangle shown in blue. The method he used is called the "quadrature of the parabola" and it is (from our modern perspective) a relatively simple though still revolutionary idea.

One very interesting consequence is that the slope of the tangent to the parabola at this midway point is equal to the slope of the chord.

The general equation of a parabola is

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

But for any given parabola, we can translate it to the origin and the parabola at the origin with the same shape is

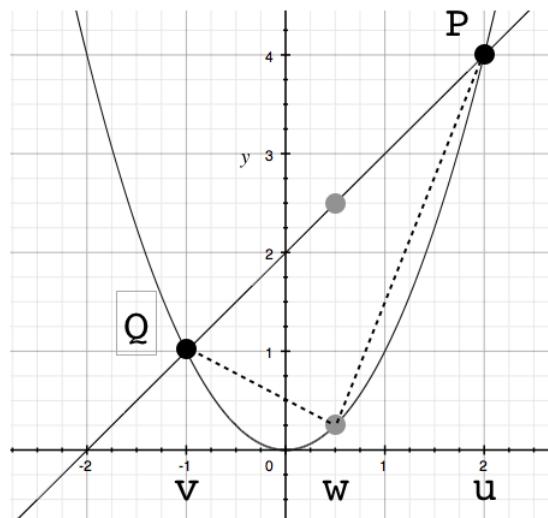
$$y = ax^2$$

This can be demonstrated by completing the square.

If we pick two points on the parabola at $x = u$ and $x = v$, then the corresponding coordinates are

$$P = (u, au^2)$$

$$Q = (v, av^2)$$



P is the right-hand point in the figure. Let us say that $au^2 > av^2$ and the slope m of the chord that connects them is

$$m = \frac{au^2 - av^2}{u - v} = \frac{a(u^2 - v^2)}{u - v} = \frac{a(u - v)(u + v)}{u - v}$$

so

$$m = a(u + v)$$

We can see that this formula gives the correct answer for $u = -v$, since the slope at the vertex is 0. Now label the midpoint $x = w$

$$w = \frac{1}{2}(u + v)$$

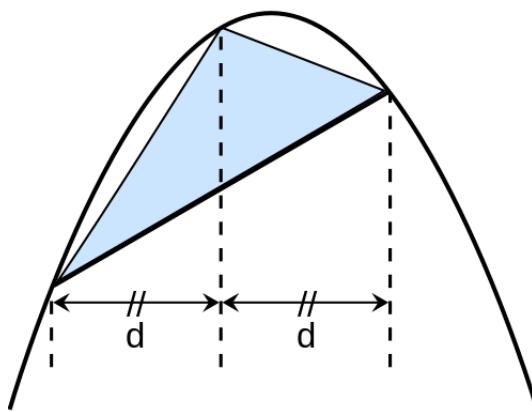
And the slope at w (from calculus) is

$$f'(w) = 2aw == 2a \frac{1}{2}(u+v) = a(u+v)$$

So the proposition is correct.

Quadrature

Another interesting thing about this figure is that the area of the triangle can be found from the length of the vertical coming down from the top.



If we simply turn the graph sideways in our mind, then the two small triangles share the part of this line within the blue region, which is their "base", b . And they both have "height" d , since w was chosen as half way between u and v , so their areas are equal, and the total area of the two together is

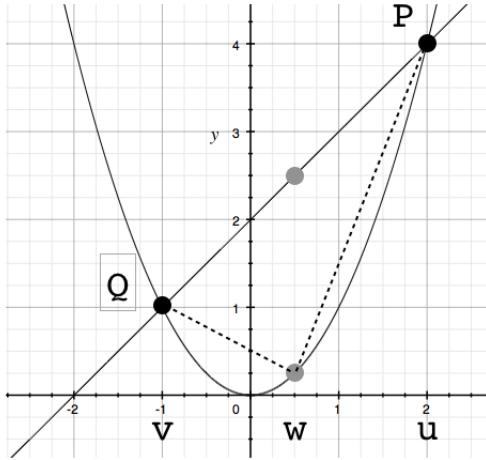
$$A = bd$$

We want to find an expression for the area only in terms of u and v (no b or d). Let's look at the second version of the figure again below.

To repeat, what we found above is that the slope at the point on the parabola corresponding to $x = w$ is equal to the slope of the line that connects v and u , and more important to us now, that the area of the combined triangle (vertices u, v, w) is

$$A = (u - w) b = \frac{1}{2}(u - v) b$$

where b is the distance parallel to the y -axis between the two points marked in gray.



The length of the "base" b is the average of the y-values for $x = u$ and $x = v$, minus aw^2 .

$$b = \frac{1}{2}(au^2 + av^2) - aw^2$$

and from before

$$w = \frac{1}{2}(u + v)$$

so we have

$$b = \frac{1}{2}(au^2 + av^2) - a \left[\frac{1}{2}(u + v) \right]^2$$

Factor out $a/4$

$$\begin{aligned} &= \frac{1}{4}a [2u^2 + 2v^2 - (u + v)^2] \\ &= \frac{1}{4}a [2u^2 + 2v^2 - u^2 - 2uv - v^2] \\ &= \frac{1}{4}a [u^2 - 2uv + v^2] \\ b &= \frac{1}{4}a (u - v)^2 \end{aligned}$$

The area is

$A = bd = \frac{1}{8}a (u - v)^3$

(36.1)

check

We'll check three cases to see if this makes sense. First if

$$u = v$$

then the area is zero and $w = u = v$, so that's good. Second, if

$$u = -v$$

then

$$A = \frac{1}{8}a(u-v)^3 = \frac{1}{8}a(2u)^3 = au^3$$

We compare this result with a direct computation by geometry. In the figure we have two symmetric triangles with individual area

$$\frac{1}{2}u \ au^2$$

The total area is twice that, so it checks. Finally, suppose we have $v = 0$

$$A = \frac{1}{8}a(u-v)^3$$

This one is harder to see, but we have that

$$d = \frac{1}{2}(u-v) = \frac{1}{2}u$$

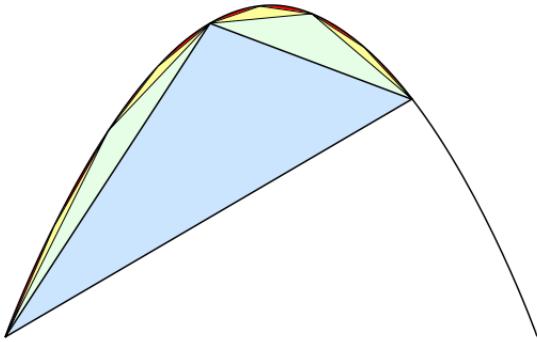
b is the distance between the average y-value which is $(1/2)au^2$ and $aw^2 = a(u/2)^2$

$$b = a\left(\frac{1}{2}u\right)^2 - \frac{1}{2}[au^2 - 0] = \frac{1}{4}a u^2$$

$$A = bd = \frac{1}{8}au^3$$

so they all check.

Quadrature of the parabola



The reason for the whole preceding argument is this. The area formula is

$$A = bd = \frac{1}{8}a(u - v)^3 = k(u - v)^3, \quad k = \text{const}$$

It is solely a function of $u - v$. Suppose we draw two new triangles (in light green, above). For each of these triangles the distance between the new vertices is one-half what we had before. So everything that we have for the big blue triangle is also true for these two new ones, just adjusted by a factor of $u' - v' = (1/2)(u - v)$.

What this means is that the area of each light green triangle is in the ratio to the blue one of $(1/2)^3 = 1/8$. But there are two of these new triangles, so the new area we added is in the ratio $1/4$.

Suppose we do it again, constructing the yellow triangles. The new area of each is in the ratio $(1/4)^3 = 1/64$ but there are now 4 of these yellow triangles so the total area is in the ratio $1/16 = (1/4)^2$

If we call the area of the original triangle T , that of the blue plus the light green is

$$A = T + \frac{1}{4}T$$

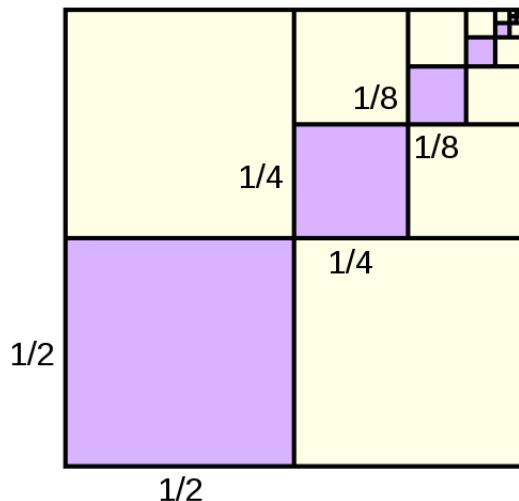
and with the addition of the yellow it is

$$A = T + \frac{1}{4}T + \frac{1}{16}T$$

so, as an infinite series it is

$$A = T\left(1 + \frac{1}{4} + \frac{1}{16} + \dots\right)$$

Here is Archimedes' proof that the sum of this series (not counting the first term) is $1/3$.



So the total is $4/3$, and the complete area under the parabola is $4/3$ the area of the triangle drawn as we described!

This called the "method of exhaustion", and not just because it's a lot of work.

Chapter 37

References

- Acheson. *The Calculus Story*.
- Acheson. *The Wonder Book of Geometry*.
- Corral. *Trigonometry*.
- Courant, Robbins, Stewart. *What is Mathematics?*.
- Densmore. *Euclid's Elements* (annotated Heath translation).
- Dunham. *Euler: The Master of Us All*.
- Dunham. *Journey Through Genius*.
- Dunham. *The Mathematical Universe*.
- Hamming. *Methods of Mathematics Applied to Calculus, Probability, and Statistics*.
- Kiselev. *Geometry, Book I. Planimetry*.
- Kline. *Calculus*.
- Lockhart. *Measurement*.
- Maor. *To Infinity and Beyond*.
- Nahin. *An Imaginary Tale. The Story of $\sqrt{-1}$* .
- Nelsen. *Proofs Without Words*.
- Posamentier, Salkind. *Challenging problems in geometry*.

- Simmons. *Precalculus mathematics in a nutshell*.
- Spivak. *The Hitchhiker's Guide to Calculus*.
- Stewart. *Significant Figures*.
- Strang. *Calculus*.
- Thompson. *Calculus Made Easy*.