A Human Introduction to Geometry **Spring 2017**
UM da Vinci Program **Drew Armstrong**

Contents

Introduction

Geometry is the most human of mathematical pursuits; so much so that it was regarded as insufficiently rigorous for twentieth century tastes and was largely banished from the undergraduate curriculum. This is a shame because drawing and looking at pictures are excellent ways to engage students. Visual intuition is also the primary way that we can hack our primate architecture, to allow us to make progress in more abstract kinds of mathematics that would otherwise be impossible to comprehend.

In this class we will embrace the human side of mathematics by following the story of geometry—pure and applied—through the ages. On the applied side we will follow geometry from its earliest use in land measurement, through the discovery of perspective drawing in the Renaissance, to its use today in computer graphics. On the pure side we will discuss

how geometry has always been close to the heart of Western science and philosophy, from Pythagoras/Plato/Euclid, through Kepler/Newton, to Kant and beyond.

At the same time, we will mix the discussion of ancient mathematics with the modern language of coordinates and transformations. Hopefully this course will provide a useful supplement and motivational examples for your other courses in math and science.

Apology

The subject of this course is historical mathematics, not mathematical history. Thus I will occasionally mix up the history with the story that we mathematicians tell ourselves about our subject. When we discuss figures like Pythagoras/Plato/Euclid I will be at least as concerned with **what their legends mean for our current understanding of mathematics** as I am for the standards of historical scholarship.

Unfortunately, the story that we mathematicians tell ourselves is currently biased towards Western sources (or, rather, Near Eastern sources that were appropriated by the West). Much work has been done recently to augment this study with sources from other cultures. In particular, China and India both had rich mathematical traditions that remained isolated from the West until modern times.[1] I will do the best I can.

1 The Pythagorean Tradition

As an entry point we begin with the birth of "capital M" Mathematics in ancient Greece. After the so-called Ionian Enlightenment in the 6th century BC,[2] a thriving school was born in which the topics of arithmetic, geometry, music and astronomy were indistinguishable. This tradition strongly influenced the history of science in the West until it was challenged by the Newtonian paradigm in the 17th century. Nevertheless, many theoretical physicists and mathematicians today still see themselves as part of the Pythagorean tradition.

1.1 Pythagoras and $\sqrt{2}$

In the earliest years of the Greek tradition **the arithmetic of whole number ratios** was regarded as the foundation of mathematics. In fact, the school of Pythagoreans regarded whole number ratios as a sort of "theory of everything". This is recorded in their famous dictum:

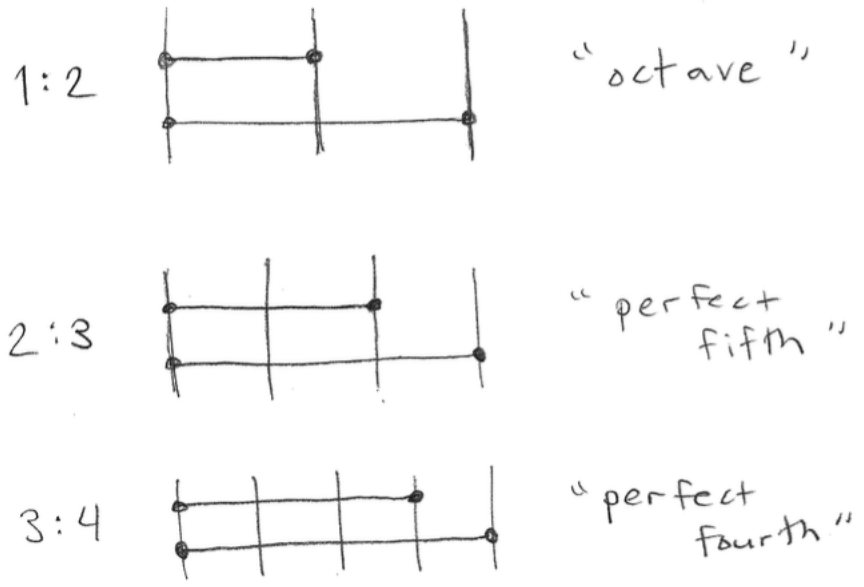
All is number.

[1] See *The Crest of the Peacock* by George Gheverghese Joseph.

[2] The 6th century BC saw the birth of important ideas all over the world, including Buddhism in India and Confucianisn and Taoism in China.

However, by the end of the classical period this dictum had been abandoned. When Euclid of Alexandria (fl. 300 BC) wrote his famous work *The Elements*, which became the standard mathematical text in the West, he conciously decided to found all of mathematics on **the geometry of lines and circles**. What happened that caused the Greeks to switch their focus from arithmetic to geometry?

The Pythagorean tradition was founded by the mysterious figure Pythagoras of Samos (c. 570–495 BC). His fundamental insight had to do with the nature of *consonance* and *dissonance* in music. Legend says that he made his discovery by listening to the sound of two blacksmith's hammers being struck simultaneously. He noticed that whether they sounded good or bad together was related to the relatize sizes of the hammers. In modern terms we would phrase the discovery as follows: two guitar strings, of the same density and under the same tension, will sound good together when their lengths are in the ratio of small whole numbers. The three most pleasing ratios are called[3] the *octave*, *perfect fifth*, and *perfect fourth*:



This discovery was so meaningful to the Pythagoreans that they founded an entire school on the idea that "all is number", where by "number" they meant *ratios of whole numbers*. (In modern terms a ratio of whole numbers is called a *rational number*). The Pythagoreans had other precepts as well, such as:

- Don't eat beans because they look like gonads.
- Be nice to dogs because they are reincarnated humans.

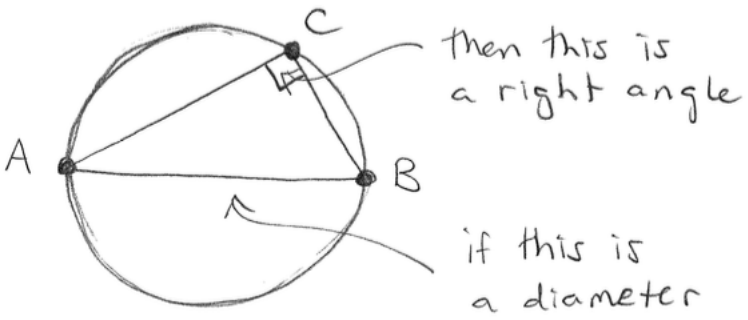
But we won't discuss those things.

[3]They are called this because they occur between the 1st, 4th, 5th and 8th notes of the modern 8 note scale.

Modern scholarship shows that sophisticated mathematics was happening in many cultures during the same period, but there is one achievement that seems unique to the Pythagorean (Ionian) tradition:

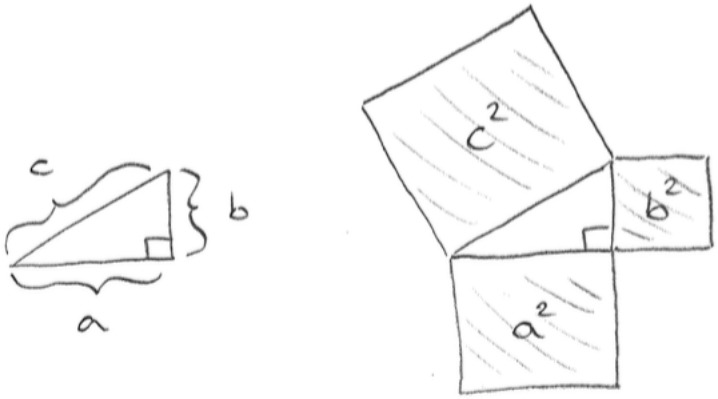
The idea of mathematical proof.

The very earliest mathematical *theorem* is attributed to Thales of Miletus (c. 625–545 BC). It says that if AB is a diameter of a circle and if C is any other point on the circle, then the line segments AC and BC make a right angle:



This fact would have been known to other cultures at the time. The peculiarity of the Ionians is that they asked the question: **why is this true?** And they came up with a satisfying answer, called a *mathematical proof*. We don't know Thales' proof[4] so instead I'll move on to a much more significant and famous result attributed to the Pythagoreans.

The Pythagoream Theorem. Consider a right triangle with side lengths a, b, c (shown on the left) and construct the squares on the three sides (shown on the right).

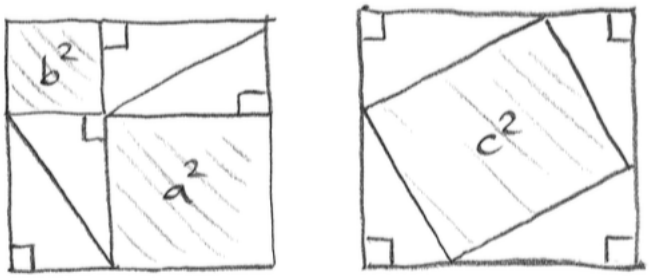


[4]In fact we know almost nothing about Thales. This theorem is probably not even his.

Then we must have $c^2 = a^2 + b^2$. In other words: the area of the square on the hypotenuse is equal to the sum of the areas of the squares on the other two sides. //

There are hundreds of ways to prove the Pythagorean Theorem. We don't know which was the original proof so I'll just show you my favorite one. The goal of the proof is to convince you[5] that the result really is true. You are encouraged to complain about the proof if you don't find it convincing.

Proof. Using the three squares and multiple copies of the original triangle we can assemble two larger squares, as in the following diagram:



Note that each of these squares has side length $a+b$ and so they must have the same area. Note also that each square contains four identical copies of the original triangle. If we **remove** these four triangles from each side then whatever remains must still be equal (equals subtracted from equals are still equal). Thus we conclude that $a^2 + b^2$ (which is the remaining area on the left) equals c^2 (which is the remaining area on the right). □

Remarks:

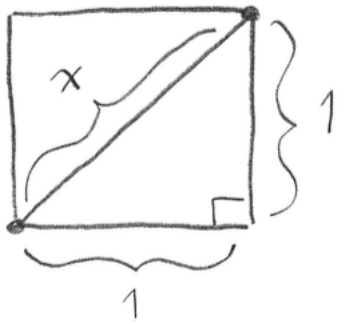
- Note that this is a completely **geometric** argument. At no point did we need to compute anything algebraic like $(a + b)^2 = a^2 + 2ab + b^2$.
- Did you find the proof convincing? If you are perceptive, you might think to ask **why** the figure on the right side is really a square. It looks like a square but how do we know that the four sides are straight lines? [There is a secret assumption hiding here that the three interior angles of a triangle add to 180°, i.e., a straight line. Why is **that** true? We will return to this issue later.]

As a simple corollary of the Pythagoream Theorem we have the following fact.

Corollary. The ratio between the side and diagonal of a square is $\sqrt{2}$. //

[5] yes, you

Proof. Consider a square with side length 1 and let x denote the length of a diagonal, as in the following diagram:



Note that the diagonal divides the square into two right triangles. Applying the Pythagorean Theorem to one of these triangles gives

$$\begin{aligned} x^2 &= 1^2 + 1^2 \\ x^2 &= 1 + 1 \\ x^2 &= 2 \\ x &= \sqrt{2}. \end{aligned}$$

□

This is a very modern kind of proof. [The use of the symbol x for an "unknown quantity" didn't become common until after the work of Descartes in 1630. More on this later.] The Greeks would have tried to express the answer in purely geometric terms. In fact, they **did** try to do this but they ran into an embarassing problem. The following result was reputedly discovered by a Pythagorean named Hippasus (fl. 5th century BC). Legend says that the result was so damaging to the Pythagorean world view that Hippasus was punished by drowning at sea. In any case, this result led to a crisis in the foundations of mathematics.

First I'll present the theorem and its proof in completely modern terms, then we'll discuss why it was so controversial.

Theorem. The square root of 2 is not a ratio of whole numbers. //

Proof. We will use a method called *proof by contradiction*. To do this we will make a hypothetical assumption and just see where it leads. Here is our hypothetical assumption:

Assume that the square root of 2 **can** be expressed as a ratio of two whole numbers.

We don't know what these whole numbers are so we can just call them a and b. Thus we have assumed that $\sqrt{2} = a/b$. Now we can square both sides of this equation to obtain

$$\begin{aligned} \sqrt{2} &= a/b \\ (\sqrt{2})^2 &= (a/b)^2 \\ 2 &= a^2/b^2 \end{aligned}$$

and then we can multiply both sides of this equation by the whole number b^2 to obtain

$$\begin{aligned} 2 &= a^2/b^2 \\ 2b^2 &= a^2. \end{aligned}$$

In particular, this equation tells us that a^2 is an **even number** because it is 2 times the whole number b^2. In this case I claim that a must also be even. Indeed, you might recall the following multiplication table for even and odd numbers:

$\times$	even	odd
even	even	even
odd	even	odd

If a were odd then $a^2 = a \times a$ would also be odd. Thus the only possibility is that a is even, which means that we can write $a = 2a'$ for some whole number a'. Now we can substitute this into the previous equation to obtain

$$\begin{aligned} 2b^2 &= a^2 \\ 2b^2 &= (2a')^2 \\ 2b^2 &= 4(a')^2 \\ 2b^2/2 &= 4(a')^2/2 \\ b^2 &= 2(a')^2. \end{aligned}$$

But now this equation tells us that b^2 is **an even number**, and by the same reasoning as before we must have $b = 2b'$ for some whole number b'. Substituting once more gives

$$\begin{aligned} b^2 &= 2(a')^2 \\ (2b')^2 &= 2(a')^2 \\ 4(b')^2 &= 2(a')^2 \\ 4(b')^2/2 &= 2(a')^2/2 \\ 2(b')^2 &= (a')^2. \end{aligned}$$

What have we done here? We began with a solution to the equation $2b^2 = a^2$ in positive whole numbers a, b and now we have produced another solution $2(b')^2 = (a')^2$ in positive whole numbers a', b'. Furthermore, these new numbers satisfy

$$\begin{aligned} a &> a' > 0 \\ b &> b' > 0. \end{aligned}$$

So far this is not very interesting. The key observation is that **we can repeat the process indefinitely**. By running the argument again and again we will obtain two infinite decreasing sequences of positive whole numbers:

$$\begin{array}{ccccccccc} a & > & a' & > & a'' & > & \cdots & > & 0 \\ b & > & b' & > & b'' & > & \cdots & > & 0 \end{array}$$

Can you imagine such a thing as an "infinite decreasing sequence of positive whole numbers"? Neither can I. Since our original assumption leads to an absurdity we conclude that it must have been wrong. That is, we conclude that it is **not** possible to express $\sqrt{2}$ as a ratio of whole numbers. □

Remarks:

- The method of proof by contradiction is also called *reductio ad absurdum*: if an assumption can be carried to an absurd extreme then the assumption must be false. The particular flavor of this argument used above is called the *method of infinite descent*: if an assumption leads to the construction of an infinite descending sequence of positive whole numbers then the assumption must be false.
- But **why** is it absurd to have an infinite descending sequence of positive whole numbers? This seems intuitively clear based on our experience of "whole numbers". That is, whole numbers are separated from each other so you can't squeeze an infinite amount of them into a finite space. In modern mathematics we have decided that this principle **can't be proved** from anything more basic, so we take it as a fundamental assumption (called an *axiom*). This axiom goes by many names, two of the most common being the *principle of induction* and the *well-ordering principle*.

Today we would say that $\sqrt{2} = 1.4142\cdots$ is an *irrational number*, but the Pythagoreans didn't think like this. To them the square root of 2 was not a "number" at all, and this was a disturbing challenge to their dictum that "all is number". Indeed, the diagonal of a square is a basic geometric contruction, so any "theory of everything" that can't handle it must not be very good theory.

The Pythagoreans would think of it this way: Let s be the side length of a square and let d be the length of the diagonal of the same square. We want to find a common unit of measure (let's call it u) to compare them. So let's assume that d is a units long and s is b units long for some **whole numbers** a and b (we do not assume that s and d are whole numbers). Thus we have $s = au$ and $d = bu$. Now the Pythagorean Theorem says that $d^2 = s^2 + s^2$ and so we must have

$$\begin{aligned} d^2 &= s^2 + s^2 \\ (au)^2 &= (bu)^2 + (bu)^2 \\ a^2u^2 &= b^2u^2 + b^2u^2 \\ a^2\cancel{u^2} &= (b^2 + b^2)\cancel{u^2} \end{aligned}$$

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

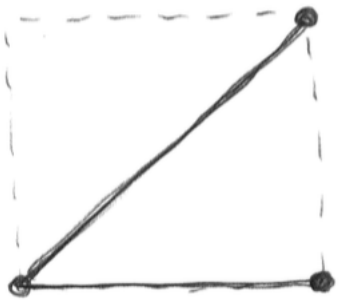
But we just finished proving that this equation of whole numbers is impossible. Therefore, we can say that

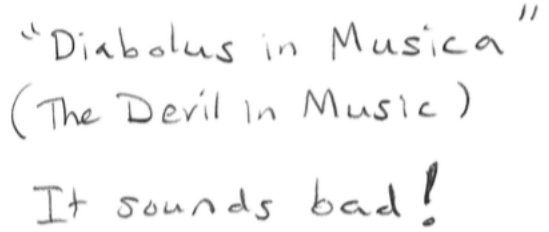
the side and diagonal of a square are incommensurable.

This came to be known as the "crisis of incommensurables", and it led the Greeks to become distrustful of the concept of "number".

Musical Remark. There is also a fun musical connection here. If we make two guitar strings on the side and the diagonal of a square (with the same density and under the same tension) then their lengths are in the ratio $1 : \sqrt{2}$. Since this is not a ratio of small whole numbers (indeed, it's not a ratio of **any** whole numbers) the Pythagorean theory of harmony tells us that the strings will sound bad when played together.

You can try this yourself. It turns out that this interval is exactly half of an octave. On the piano it is any interval of six semitones (for example from C to F#). This interval does indeed "sound bad" and it is almost never used in a melody. Today this interval is know as the "tritone" but during the Middle Ages it was called the *Diabolus in Musica* (the Devil in Music). Legend says that it was against the law to play this interval in church.





1.2 Plato and Regular Polyhedra

It is doubtful that Pythagoras himself is responsible for any of the mathematics just discussed, but it is certainly true that there was an active group of Pythagorean mathematicians in the fifth century BC. The Pythagoreans also became significantly associated with cosmology and with a theory called

the harmony of the spheres.

This idea says that the Earth is at the center of a nested family of transparent crystal spheres. Each planet (*asteres planetei* literally means "wandering star") is attached to its own sphere and these spheres rotate with respect to each other. The fixed stars of the constellations

are contained on the outermost sphere.[6] We know that crystal spheres rubbing together will create a sound and so there was some speculation as to what sound the universe makes.

Of course the sound of the universe should be harmonious so it was assumed that the motion and sizes of the celestial spheres should be governed by small whole number ratios. Some of the earliest musical scales are based on the hypothetical "cosmic scale". In fact, our modern major scale is attributed to Claudius Ptolemy (c. 100–170 AD).[7] This is the same Ptolemy whose theory of planetary motion became standard dogma in the Middle Ages and persisted until the time of Johannes Kepler (1571–1630 AD).

Theories of celestial harmony became highly developed in the Pythagorean tradition but our knowledge of this comes from very few sources. The earliest complete work on the subject is Plato's dialogue *Timaeus*, which we will discuss in this section.

Plato (c. 427–347 BC) dominated Greek philosophy during his time. He is famous as the founder of the Academy in Athens, which became the archetype for institutes of higher education in the West. Plato was also very much a Pythagorean in his views (as opposed to his student Aristotle whose inclinations were less mathematical). Legend says that the entrance to the Academy bore the inscription:

Let no one ignorant of geometry enter here.

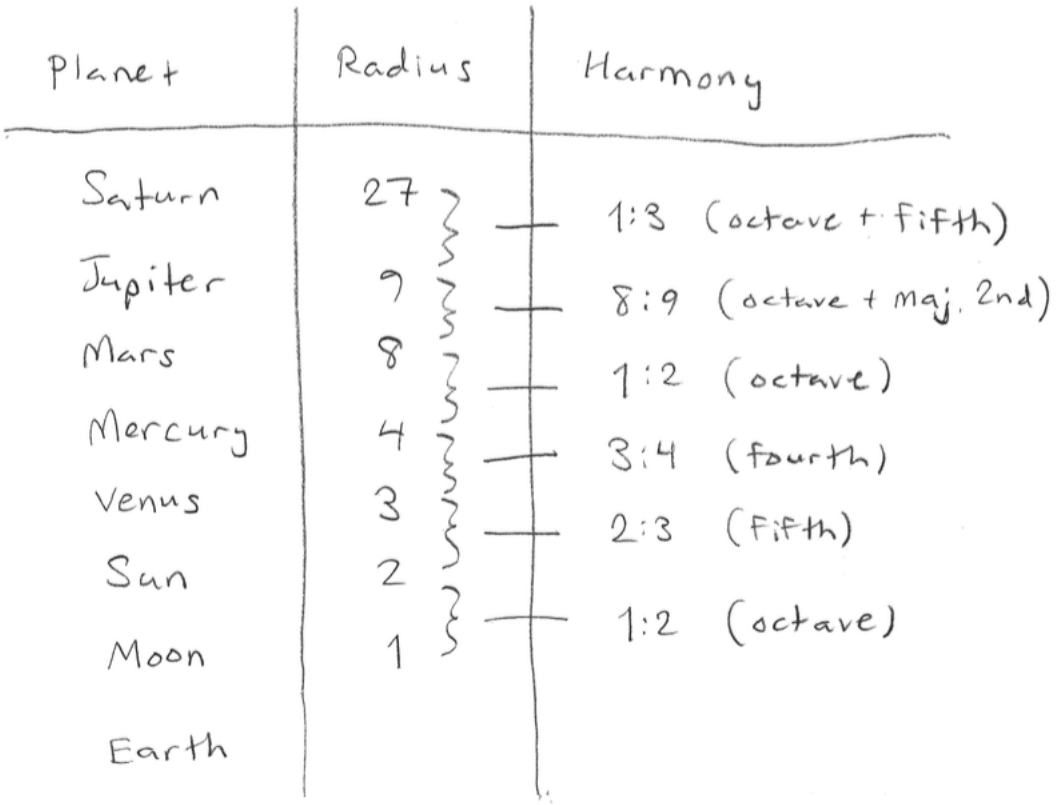
We know a lot about Plato because almost all of his work survived into modern times, even though it was temporarily lost in the West. It was preserved in the Middle East and then recovered in the West during the Renaissance. Plato's dialogue *Timaeus* (c. 360 BC) is his only work that was never lost in the West and as such it has exerted a disproportionate influence on European thought. The dialogue is mostly a monologue given by the character Timaeus of Locri who is a philosopher of the Pythagorean school. Timaeus describes elaborate theories of all aspects of the physical world. For current purposes I will only mention his theories of

- planetary motion,
- atomic physics.

In Timaeus' theory of planetary motion the 7 known planets are attached to 7 rotating crystal spheres, all centered on the Earth (which was known to be a sphere). Timaeus lists the order of the planets and their relative radii as follows:

[6] Possibly they are holes punched in an opaque sphere with their light coming from outside.

[7] It is known as Ptolemy's intense diatonic scale. `https://en.wikipedia.org/wiki/Ptolemy's_intense_diatonic_scale`



The association between planetary radii and musical harmony is left vague in the *Timaeus*. It seems that all of the constructions are based on the sequence of powers of 2 and 3:

$$1, 2, 3, 4(=2^2), 8(=2^3), 9(=3^2), 16(=2^4), 27(=3^3), \text{ etc.}$$

This sequence of numbers was the foundation of Pythagorean musical theory. We can use this to give a modern explanation[8] for the fact that musical scales have 7 different notes. The explanation goes as follows: The two most important musical intervals are 1:2 (the octave) and 2:3 (the perfect fifth). To compare them we need to find whole numbers m and n so that

$$m \text{ octaves } = n \text{ perfect fifths.}$$

This problem is analogous to the problem of the commensurability of the side and diagonal of a square, and in this case we will also find that the problem is impossible. However, the proof is much easier in this case.

Theorem. The octave and the perfect fifth are incommensurable. //

Proof. Since musical intervals are **ratios**, they are combined by **multiplication**. In mathematical terms we want to find whole numbers m and n satisfying the following equation:

$$m \text{ octaves } = n \text{ perfect fifths}$$

[8] i.e., not related to the fact that there are 7 planets

$$\underbrace{\frac{1}{2} \times \frac{1}{2} \times \cdots \times \frac{1}{2}}_{m \text{ times}} = \underbrace{\frac{2}{3} \times \frac{2}{3} \times \cdots \times \frac{2}{3}}_{n \text{ times}}$$

$$\left(\frac{1}{2}\right)^m = \left(\frac{2}{3}\right)^n$$

$$\frac{1}{2^m} = \frac{2^n}{3^n}$$

$$3^n = 2^n \times 2^m$$

$$3^n = 2^{m+n}.$$

But observe that this equation is **impossible** because 3^n is always an **odd** number (odd times odd is odd) and 2^{m+n} is always an **even** number (even times even is even). □

But music must proceed anyway, so instead of an exact solution we are willing to accept a good approximate solution. It turns out that the approximation

$$3^{12} \approx 2^{7+12}$$

is good enough. In other words, we have

$$7 \text{ octaves } \approx \text{ 12 perfect fifths.}$$

This is the reason that our musical scales have 7 notes and the reason that we divide the octave into 12 equal semitones. What I mean by "good enough" is that this approximation is close enough to perfect Pythagorean harmony that the human ear usually can't tell the difference.

So much for Timaeus' theory of the very large. Next we turn to his theory of the very small. To describe this theory we first need to discuss the concept of *regular polyhedra*.

Definition of Regular Polyhedra. A polyhedron is a convex three-dimensional shape built from a finite number of vertices, edges and faces (*polyhedron* literally means "many-sided"). We say that such a polyhedron is *regular* if each vertex/edge/face looks the same as every other vertex/edge/face. In other words,

- each vertex meets the same number of faces,
- each edge has the same length,
- each face has the same number of vertices.

The following classification theorem is attributed to the Pythagorean mathematician Theaetetus of Athens (c. 417–368). The earliest surviving reference to the result is Plato's *Timaeus*; for this reason the regular polyhedra are also known as *Platonic solids*. Examples of Platonic

solids were known in other cultures.[9] Again, the Greek contribution was to recognize that these shapes can be logically classified; it turns out that there are exactly five of them.

Theorem. There are five regular polyhedra:

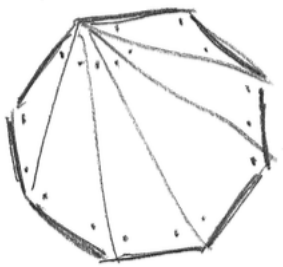
tetrahedron, cube, octahedron, dodecahedron, icosahedron.

//

The proof will require the following preliminary result (called a *lemma*).

Lemma. Each interior angle of a regular n-gon is equal to $\frac{n-2}{n}$ times 180°. //

Proof of the Lemma. Consider a regular n-gon and divide it into triangles. Note that there will be a total of $(n-2)$ triangles in the decomposition. For example, the following figure shows that an octagon ($n = 8$) decomposes into 6 triangles:



Since the interior angles of each triangle sum to 180°, we see that the sum of all interior angles in the n-gon is $(n-2) \times 180^\circ$. Now erase the triangles. Since each of the n interior angles of the n-gon is the same, we conclude that each of them must be equal to

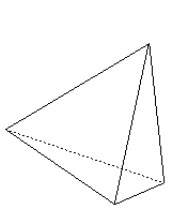
$$\frac{(n-2) \times 180^\circ}{n} = \left(\frac{n-2}{n}\right) \times 180^\circ.$$

□

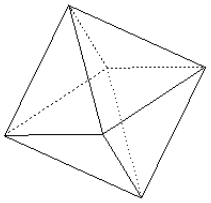
Proof of the Theorem. We will prove that there are only five regular polyhedra. So let P be a regular polyhedron. Suppose that each of its faces has n vertices. Since all edges have the same length, each face is a regular n-gon. We also know that exactly d of these faces meet at each vertex. Our goal is to find restrictions on the numbers n and d.

Case $n = 3$. Assume that each face is an equilateral triangle (i.e., with interior angles 60°). Now what are the possibilities for d? That is, how many equaliteral triangles can meet at a vertex? For $d = 3, 4, 5$ we obtain the regular *tetrahedron*, *octahedron*, and *icosahedron*, as shown here:

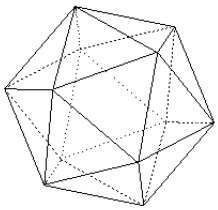
[9] For example, carved stone versions dating from 2000 BC have been found in Scotland.



Tetrahedron

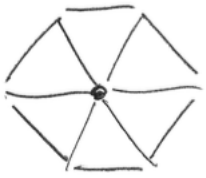


Octahedron



Icosahedron

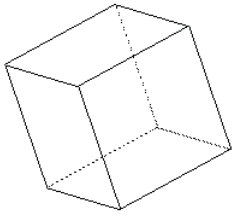
If we try to place $d = 6$ equilateral triangles around a vertex then since $6 \times 60^\circ = 360^\circ$ the vertex flattens out and stops being a vertex at all:



6 triangles around a vertex is flat.

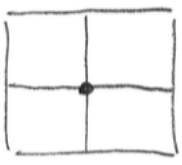
For $d \geqslant 7$ we have $d \times 60^\circ > 360^\circ$ and so it is impossible to fit d equilateral triangles around a vertex without bending them.

Case $n = 4$. Assume that each face is a square (i.e., with interior angles 90°). Now what are the possibilities for d? That is, how many squares can meet at a vertex? For $d = 3$ we obtain the *cube* (or regular *hexahedron*):



Hexahedron

If we try to place $d = 4$ squares around a vertex then since $4 \times 90^\circ = 360^\circ$ the vertex flattens out and stops being a vertex at all:



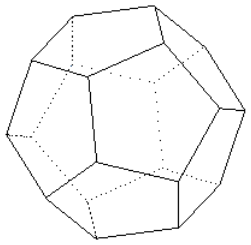
4 squares around a vertex is flat.

For $d \geqslant 5$ we have $d \times 90^\circ > 360^\circ$ and so it is impossible to fit d squares around a vertex without bending them.

Case $n = 5$. Assume that each face is a regular pentagon. From the above lemma we know that each interior angle of a regular pentagon is equal to

$$\left(\frac{n-2}{n}\right) \times 180^\circ = \frac{3}{5} \times 180^\circ = 108^\circ.$$

Now what are the possibilities for d? That is, how many pentagons can meet at a vertex? For $d = 3$ we obtain the regular *dodecahedron*:



dodecahedron

For $d \geqslant 4$ we have $d \times 108^\circ > 360^\circ$ and so it is impossible to fit d regular pentagons around a vertex without bending them.

Case $n = 6$. Assume that each face is a regular hexagon and note that each interior angle of a regular hexagon is equal to

$$\left(\frac{n-2}{n}\right) \times 180^\circ = \frac{4}{6} \times 180^\circ = 120^\circ.$$

If we try to place $d = 3$ hexagons around a vertex then since $3 \times 120^\circ = 360^\circ$ the vertex flattens out and stops being a vertex at all:

3 hexagons around a vertex is flat.

For $d \geqslant 4$ we have $d \times 120^\circ > 360^\circ$ and so it is impossible to fit d regular hexagons around a vertex without bending them.

Case $n \geqslant 7$. Each face is a regular n-gon with interior angles $(n-2)/n \times 180^\circ$. If $n \geqslant 7$ then we have $(n-2)/n \geqslant 5/7$ and hence

$$\left(\frac{n-2}{n}\right) \times 180^\circ \geqslant \frac{5}{7} \times 180^\circ \approx 128.57^\circ.$$

Then for any $d \geqslant 3$ we have

$$d \times \left(\frac{n-2}{n}\right) \times 180^\circ \geqslant 3 \times 128.57^\circ > 360^\circ$$

and hence it is impossible to place any number of n-gons around a vertex without bending them. This completes the proof. □

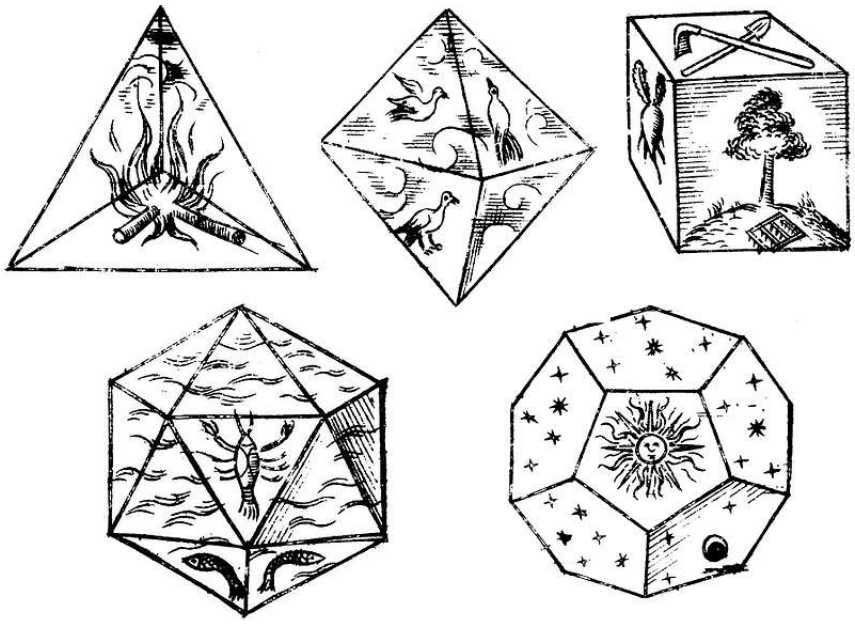
The fact that there are only five regular polyhedra seems to give them special significance. In the *Timaeus*, Plato uses them in a surprising way for his theory of atomic physics.

The first atomic theory is attributed to the Ionian philosopher Democritus (c. 460–370 BC). It states that all physical things is composed of *atoms*. These are tiny and indivisible particles (*atomos* literally means "un-cuttable") and between them is empty space. On the other hand, the Ionian philosopher Empedocles (c. 490–430 BC) is known for the theory that all physical things are composed from four original substances: Earth, Air, Fire, and Water. In the *Timaeus*, Plato combined these two theories in a creative and beautiful way. The character Timaeus states that all physical things are composed of atoms and that these atoms come in exactly five kinds: Earth, Air, Fire, Water, and Aether. Timaeus uses the word "elements" (*stoicheia*) for the different kinds.

And why are there five elements? Because there are five regular polyedra. Timaeus claims that the atoms of the various elements come in the shape of tiny regular polyedra as follows:

Element	Shape of the Atoms
Fire	tetrahedra
Air	octahedra
Water	icosahedra
Earth	cubes
Aether	dodecahedra

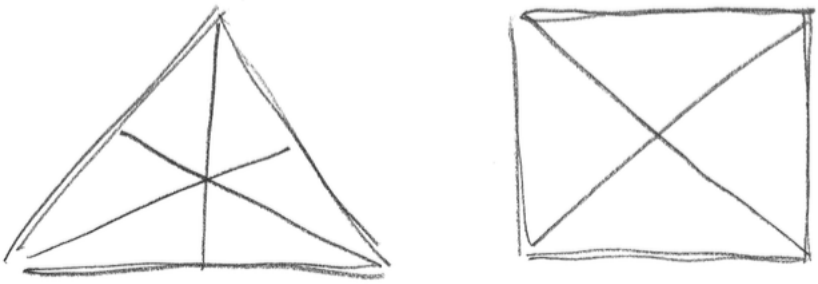
Here is a beautiful engraving of the five Platonic elements taken from Johannes Kepler's[10] *Harmonices Mundi* (1619):



It seems that the inclusion of a fifth element was necessary for mathematical reasons. Plato gets around this by saying that everything on earth is composed of Earth, Air, Fire, and Water, while Aether is used for the sphere of constellations. Thus the dodecahedron is somehow associated with the outermost celestial sphere, uniting the theories of the very small and the very large.

But that's not all. In Plato's *Timaeus* the atoms are not literally indivisible. Each atom can be further broken down into two-dimensional triangles by subdividing its faces. Each triangular face is subdivided into six $30^\circ/60^\circ/90^\circ$ triangles and each square face is subdivided into four $45^\circ/45^\circ/90^\circ$ triangles. (He does not say how the dodecahedron decomposes.) We will call these triangles Type I and Type II:

[10] We'll discuss his role in the next section.



The following table counts the number of triangles of Types I and II in a single atom of each element:

	Type I	Type II
Fire	24	0
Air	48	0
Water	120	0
Earth	0	24

From this we see that the element Earth can not be transmuted into any other element, but the elements Fire/Air/Water can be rearranged in various ways. For example, since there are 120 Type I triangles in a single atom of Water and since $120 = 2 \cdot 48 + 1 \cdot 24$, this atom can be broken down and then reassembled into two atoms of Air and one atom of Fire. We obtain the following chemical formula:

$$1W = 2A + 1F.$$

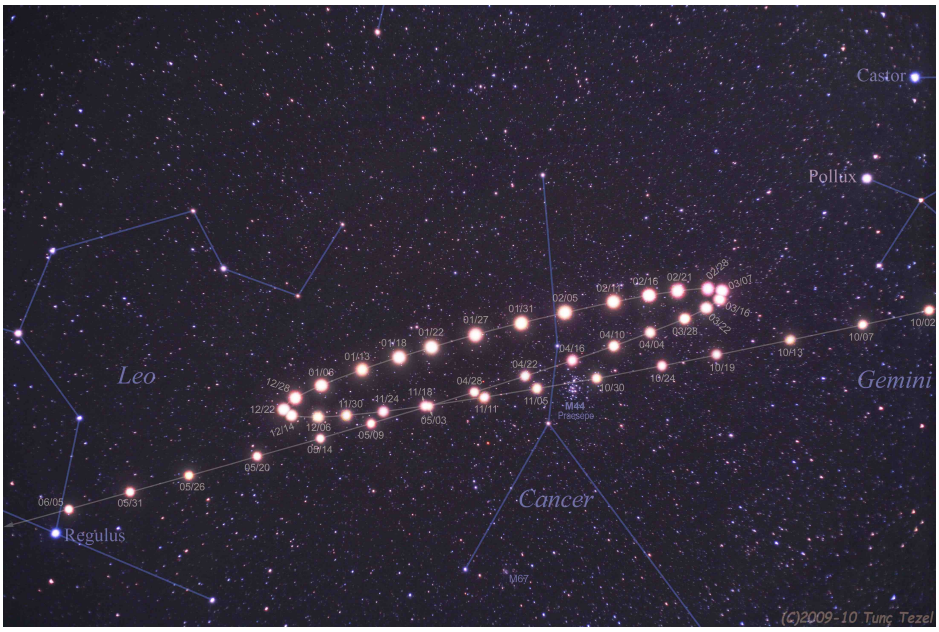
There is a surprising resonance here with the modern formula $1W = H_2O = 2H + 1O$.

1.3 Kepler and Conic Sections

Plato's *Timaeus* is a remarkable example of human ingenuity applied to the problems of physical science in the absence of sufficient experimental data. Most of Plato's theories were very far from being testable in his time, thus they had to stand or fall on aesthetic grounds.

Plato also differed with his student Aristotle on whether reason or experience should take priorty: if a theory was *beautiful*, Plato took this as evidence that it must also be *true*.[11] The key feature of the Pythagorean tradition is that it finds this beauty in the realm of mathematics and geometry.

It was not until the twentieth century that technology was able probe the atomic nature of matter. However, there were plenty of observations of the heavens and plenty of competing theories to explain them. One of the main problems that any theory had to deal with was the *retrograde motions* of the planets. If we consider the furthest sphere of constellations as fixed, then each planet moves across this background. Occasionally as the planet travels it will pause, back up, and then continue on its way. (This is why the planets are literally "wandering stars".) This phenomenon is particularly apparent with Mars and it was observed by Egyptian astronomers in the 2nd millennium BC. Here is a series of photographs taken by amateur astronomer Tunc Tezel in 2009–2010 showing the retrograde motion of Mars:



Plato's *Timaeus* was vague on this issue and there was a vigorous development of alternative theories. The key problem was to explain the retrograde motion of planets while preserving the perfection of **uniform circular motion**. The Neo-Platonic philosopher Simplicius of Cilicia (c. 490–560 AD) attributed the problem to Plato as follows:

> Plato lays down the principle that the heavenly bodies' motion is circular, uniform, and constantly regular. Thereupon he sets the mathematicians the following problem: what circular motions, uniform and perfectly regular, are to be admitted

[11] We have a similar situation today with "string theory" in physics. This is a beautiful mathematical/geometric theory of reality that is unfortunately not testable with today's technology.

> as hypotheses so that it might be possible to save the appearances presented by the planets?

That is, it was necessary to

save the appearances

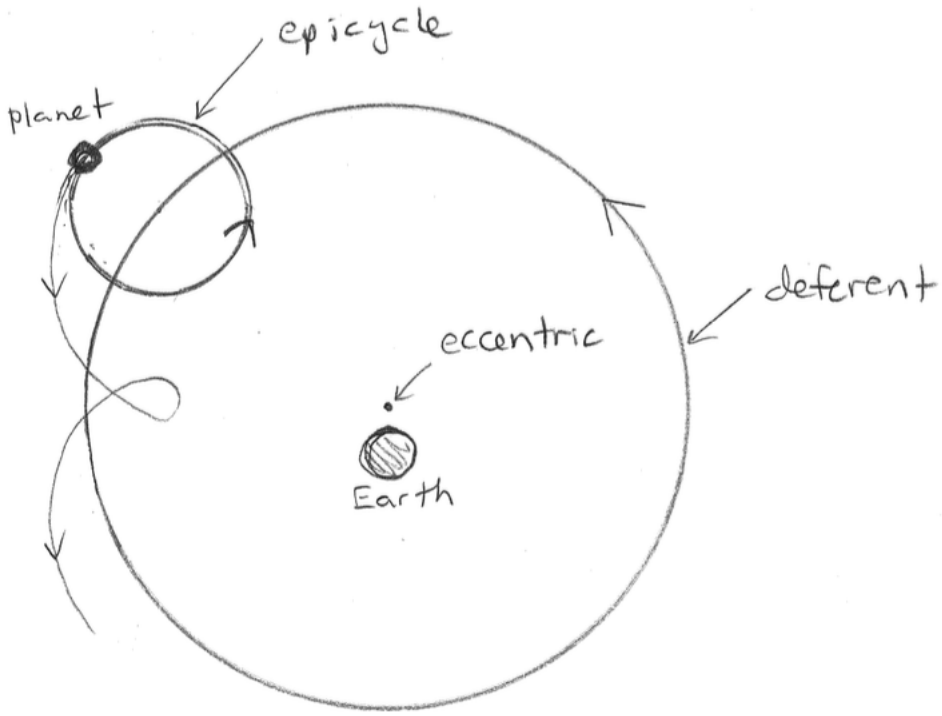
of planetary motion—which were apprently imperfect—with some underlying mathematical perfection.

The most refined astronomical theory of the Classical world was presented by Claudius Ptolemy (c. 100–170 AD) in his work the *Almagest*.[12] In Ptolemy's theory the spherical Earth is at the center of a spherical universe. The Earth is stationary while the heavens and planets rotate. As with Plato's theory, each planet is associated with a transparent crystal sphere. Here is a visualization of Ptolemy's universe taken from Peter Apian's *Cosmographia* (1524):

To explain the retrograde motion of the planets, Ptolemy adopted the idea of *deferents* and *epicycles* from Apollonius of Perga (c 262–190 BC). The "deferent" is the primary sphere of the planet; to account for astronomical observations Ptolemy sometimes allowed a deferent to be centered on an imaginary point outside the Earth, called the *eccentric*. The "epicycle" is a much smaller crystal sphere which has its center fixed at some point on the deferent. The deferent and epicycle both undero uniform cirular motion but they are allowed to rotate at different speeds. The rotation of the deferent explains the broad motion of the planet across

[12] This work was originally called *Mathematike Syntaxis* (Mathematical Treatise) in Greek, which later became the *Magna Syntaxis* (The Great Treatise) and was translated into Arabic as *al-majisti*. The title *Almagest* was adopted when the work was translated from Arabic back into Latin in the 12th century.

the constellations and the rotation of the epicycle accounts for retrograde motion.



Ptolemy's model of deferents and epicycles turned out to be extremely accurate and it was eventually adopted as dogma by the Catholic church. As astronomical observations improved, later astronomers found they could "save the appearances" by adding extra epicycles (epi-epicycles) into the theory. From a modern mathematical point of view we now realize that Ptolemy's model can be made **arbitrarily accurate** by adding epicycles upon epicycles upon epicycles.[13]

To keep up with new observations the Ptolemaic theory became more and more complicated. The original goal of the theory was to explain astronomical observations while at the same time preserving the Platonic ideals of:

- beauty
- necessity
- simplicity

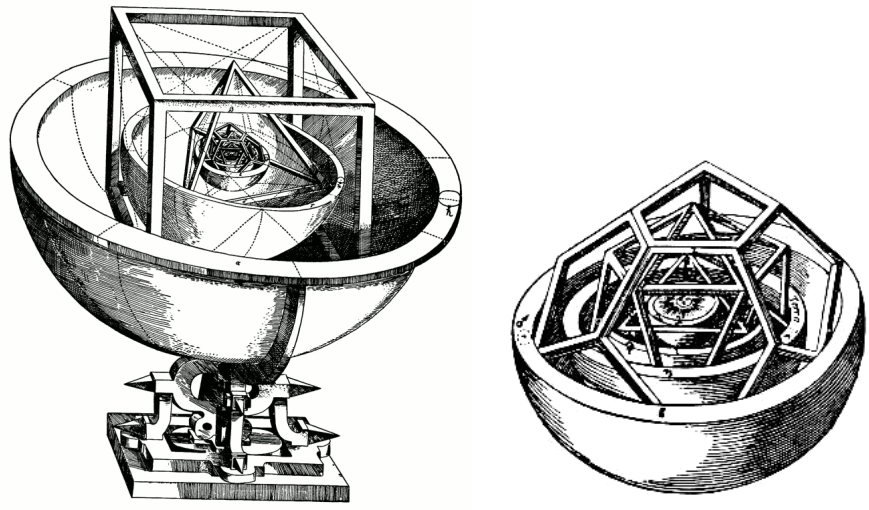
By the time of Nicolaus Copernicus (1473–1543) the Ptolemaic model had become bloated with epicycles and arbitrarily chosen parameters. This was also a time when ancient Greek writings were becoming available in Europe. Following the ancient heliocentric theory of Aristarchus

[13]This is the modern theory of "Fourier series". As a desmonstration see the YouTube video called "Ptolemy and Homer".

(c. 310–230 BC), Copernicus found that he could reduce the complexity of the Ptolemaic system by placing the Sun at the center of the universe and allowing the Earth to move. The most dramatic improvement of the heliocentric theory is that it **automatically explains retrograde motion**: it is caused when the Earth laps another planet on their mutual trip around the sun. But Copernicus' goal was not necessarily to improve the **accuracy** of the model; in fact his model was **less** accurate than the best Ptolemaic models at the time. Instead he was motivated by the Platonic ideals outlined above, which he also expressed in the language of Christian theology. In any case, Copernicus' contribution was to weaken the Ptolemaic dogma, which allowed other astronomers to experiment with new ideas.

One of those who experimented with new models of planetary motion was Johannes Kepler (1571–1630). Kepler is a fascinating figure in the history of science because his work is the bridge between two different eras: the classical Pythagorean and the modern Newtonian.

On one hand, Kepler was very much a Pythagorean, as demonstrated by his early work the *Mysterium Cosmographicum*[14] (1596). Ostensibly this is a work about planetary motion, but it could easily be mistaken for a work of pure geometry or even music theory. The work boldly defends Copernicus' heliocentric idea and it goes even further by abandoning the deferents and epicycles of Ptolemy. In their place Kepler describes a solar system built around the five Platonic solids. He imagines the solar system as a series of **6 concentric planetary spheres with the 5 Platonic solids wedged tightly between them**. To Kepler this is a very neat and compelling explanation for the fact that there are exactly 6 planets. Here is a picture of Kepler's solar system together with a close-up of the sphere of Mars:



The following table lists the planetary spheres in order, together with the Platonic solids between them and the ratios of their radii, which are computed geometrically. The final column shows the modern approximate values for the ratios. Note that the agreement is

[14]Here is the full title: *Forerunner of the Cosmological Essays, Which Contains the Secret of the Universe; on the Marvelous Proportion of the Celestial Spheres, and on the True and Particular Causes of the Number, Magnitude, and Periodic Motions of the Heavens; Established by Means of the Five Regular Geometric Solids*

pretty good:

Sphere	Shape Between Spheres	Ratio of Sphere Radii	Modern Value
Saturn	cube	$\sqrt{3} \approx 1.73$	1.73
Jupiter	tetrahedron	3	3.42
Mars	dodecahedron	$\sqrt{15-6\sqrt{5}} \approx 1.26$	1.52
Earth	icosahedron	$\sqrt{15-6\sqrt{5}} \approx 1.26$	1.38
Venus	octahedron	$\sqrt{3} \approx 1.73$	1.87
Mercury			
Sun			

On the other hand, Kepler had access to more and better astronomical data than any of his predecessors. Tycho Brahe (1546–1601) was a Danish nobleman and astronomer who spent 30 years on the island of Hven making meticulous astronomical observations. Tycho devised his own instruments (this was before the invention of the telescope in 1608) and it is said that his observervations were five times more accurate than the best available at the time. Kepler and Tycho began a correspondence in which Tycho criticized the accuracy Kepler's system from the *Mysterium Cosmographicum*. Kepler wanted to see Tycho's data, and so in 1600 he accepted an invitation to Prague to assist Tycho while a new observatory was built. When Tycho died unexpected in 1601, Kepler somehow gained access to Tycho's data and replaced him as imperial mathematician to Rudolf II of Austria.[15]

Once Kepler possessed the data he realized that his polyhedral model of the solar system was wrong and he quickly abandoned it. Then he set about trying to find a new model that would agree with the new data while still preserving the ideals of Platonic beauty. Eventually the data forced Kepler to **abandon the concept of uniform circular motion** but he found that the universe is still geometric.

[15]This is the short version of the story. I encourage you to look up the details; it's pretty interesting.

Kepler's Laws of Planetary Motion (1609–1619).

(1) The orbit of a planet is an *ellipse* with the Sun at one of the two *foci*.

(2) The line segment joining a planet to the Sun sweeps out equal areas in equal times.

(3) The square of the orbital period of a planet is proportional to the cube of the semi-major axis of its orbit.

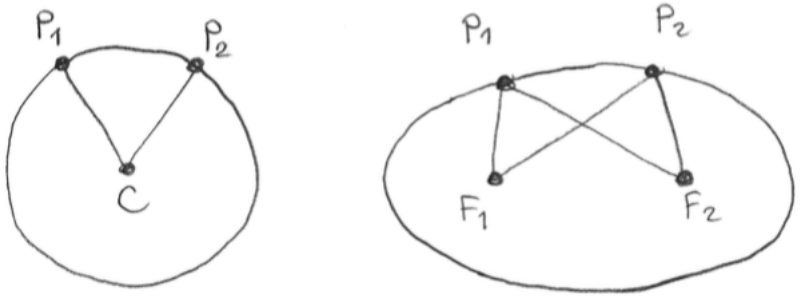
//

These laws led within a short time to the modern Newtonian view of mechanics and it turned out that they were correct enough to send 12 humans to the moon. Thus Kepler's laws are part of our modern world view; and yet Kepler still thought of his work in Pythagorean terms. He announced the Third Law in his book the *Harmonices Mundi* (Harmony of the World) which was a response to Ptolemy's work the *Harmonics*. Here Kepler argued that his model of elliptical planetary orbits is still just as perfect as Ptolemy's model in its relation to musical harmony.[16] But this was the end of an era; Kepler was the last astronomer to express his work in musical terms.

Let's examine the mathematics behind Kepler's laws. The first question is:

What is an ellipse?

To answer this we should compare the definition of an ellipse with the definition of a circle. Consider the following figure:



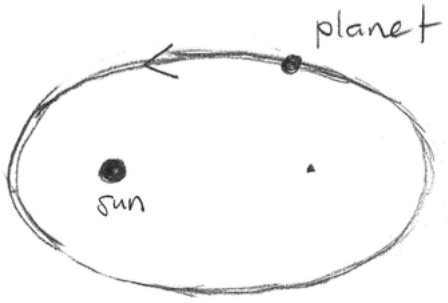
Recall that a circle is defined as the collection of all points P that are equidistant from a given point P, called the *center* of the circle. In the figure this means that the distances CP_1 and CP_2 are the same for any points P_1 and P_2. This common distance is called the *radius* of the circle. Now consider the figure on the right. The two points F_1 and F_2 are called the *foci* of the ellipse (plural of *focus*). The defining property is that for any two points P_1 and P_2 on the ellipse we must have

$$F_1P_1 + F_2P_1 = F_1P_2 + F_2P_2.$$

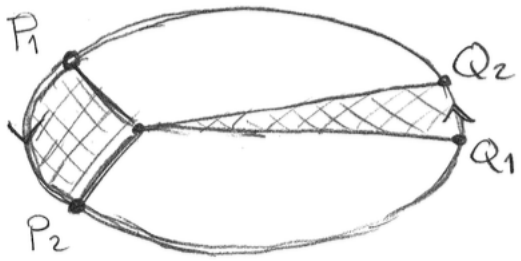
[16] Listen here: `https://youtu.be/WihmsRinpQU`

That is, an ellipse is the collection of all points P such that the sum of the distances $F_1P + F_2P$ is a constant (there is no standard name for this constant). If the two foci F_1 and F_2 approach each other then this definition degenerates into the definition of a circle. [Why?] Thus an ellilpse is a *generalization* of a circle.

Kepler's 1st Law says that each planet follows a path in the shape of an ellipse with the Sun at one of the foci:

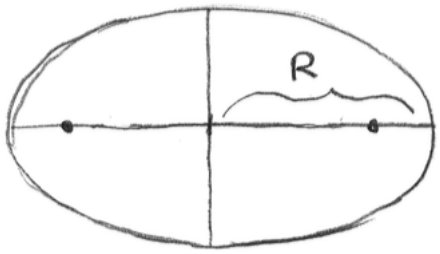


The other focus is just an imaginary point with no physical significance. This explains the behavior of the planet in **space**. To explain the behavior of the planet in **time**, Kepler needed to come up with a substitute for uniform circular motion. From the data he knew that a planet does not travel with constant speed; in fact it travels faster when it is close to the sun and slower when it is farther away. But he was still able to find a **geometric uniformity** in the motion. His 2nd Law says that planet sweeps out equal areas in equal times:



In the figure we assume that the two shaded regions have equal area. In this case, Kepler says that the planet takes the same time to travel from P_1 to P_2 as it does to travel from Q_1 to Q_2. This is a beautiful geometric explanation for the planet's changing speed.

The straight line connecting the two foci is called the *major axis* of the ellipse. I will use R to denote half of the length, since this is a generalization of the "radius" of a circle:



If we let T denote the orbital period (the amount of time it takes for one trip around the sun) then Kepler's 3rd Law says that T^2 and R^3 are *proportional*:

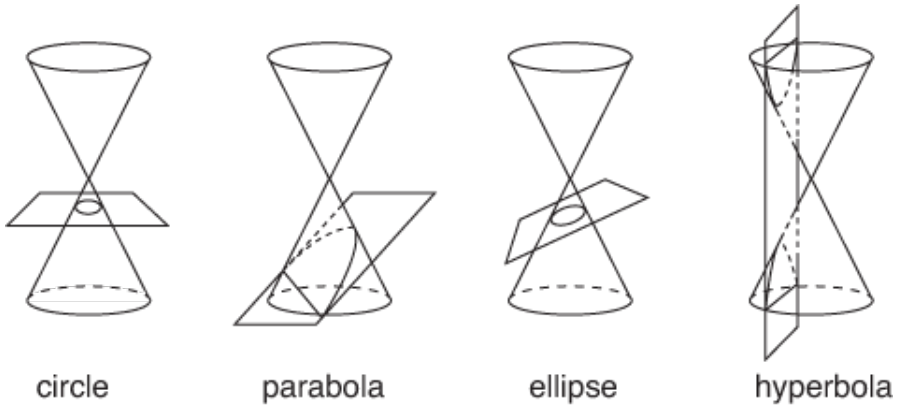
$$T^2 \sim R^3.$$

The word "proportional" means the following: If you multiply the length R by a factor of n^3 for some n then the time T will increase by a factor of n^2. Kepler called this the *Harmonic Law* because of the precise ratio between the numbers 2 and 3, and he thought of the law in explicitly musical terms.

Thus we see that Kepler's universe has just as much geometric and arithmetic beauty as the Platonic universe of circles. It is worth remarking that the geometric study of ellipses also goes back to ancient Greece: they were studied systematically in the *Conics* of Apollonius of Perga (c. 262–190 BC). The title of this work leads to the question:

What do ellipses have to do with cones?

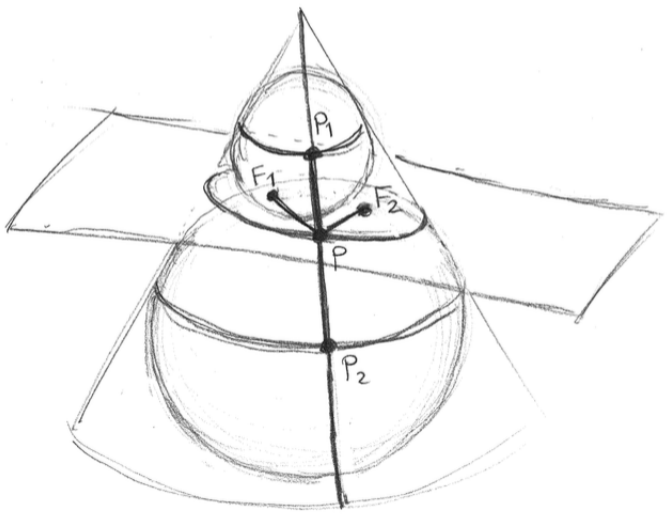
Apollonius gave us the names for *ellipses*, *parabolas*, and *hyperbolas*. Together he referred to these three kinds of shapes as *conic sections* because they can all be realized as the intersection of a flat plane with a circular cone, as in the following figure:



Finally, I will try to convince you that the definition of an ellipse in terms of its foci is equivalent to its definition as a conic section. The following beautiful proof was discovered by the Belgian mathematician Germinal Pierre Dandelin in 1822.

Theorem. If a circular cone and a flat plane have a bounded (finite) intersection, then this intersection is an ellipse. //

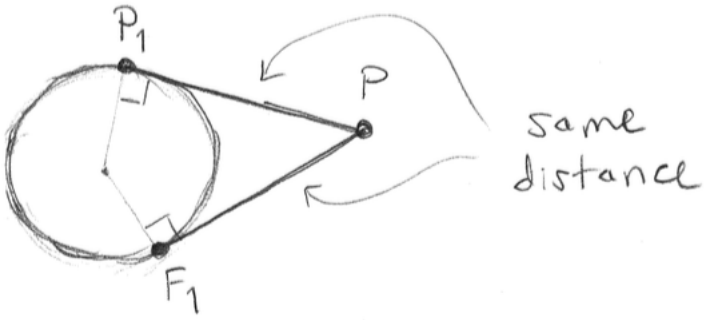
Proof. Consider a circular cone sliced by a flat plane. The intersection divides the interior of the cone into two open spaces. Now we consider the largest spheres that can fit inside each of the two spaces while touching the plane. These are called the *Dandelin spheres*:



Suppose the first sphere touches the plane at point F_1 and the second sphere touches the plane at point F_2. I claim that the intersection of the plane and cone is an ellipse with F_1 and F_2 as its foci. To prove this, consider any point P on the boundary of the supposed ellipse. We will show that the sum of the distances

$$F_1P + F_2P$$

is independent of the choice of P. Now observe that each Dandelin sphere intersects the boundary of the cone in a circle (because it is a circular cone). Now consider the unique line that connects P to the apex of the cone and let P_1 and P_2 be its points of intersection with the two circles just described. In this case I claim that $P_1P = F_1P$ and $P_2P = F_2P$. Indeed, note that the line segments P_1P and F_1P are both tangent to the first Dandelin sphere, so they must have the same length:



The same argument applied to the second sphere shows that $P_2P = F_2P$. Thus we have

$$F_1P + F_2P = P_1P + P_2P = P_1P_2.$$

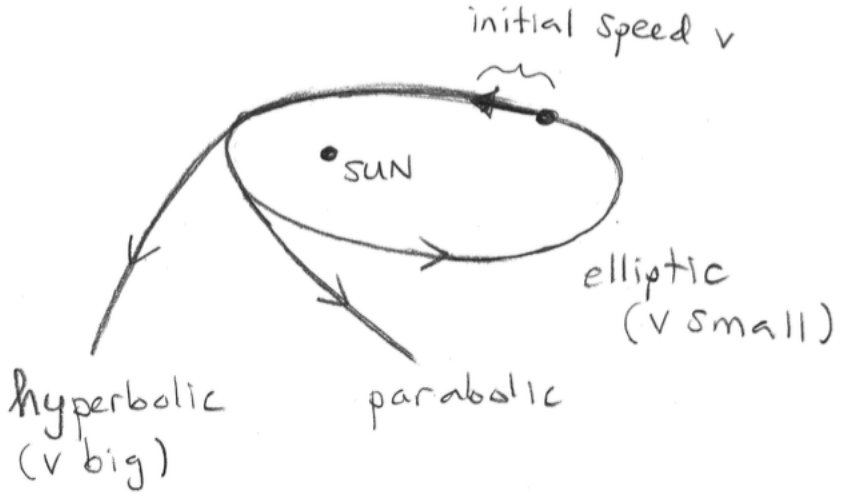
The equality on the right follows from the fact that the points P, P_1 and P_2 all lie on the same line. We conclude that the sum of the lengths $F_1P + F_2P$ is equal to the length of the line segment P_1P_2. But observe that **the length P_1P_2 is independent of the choice of P** because the two circles are parallel. We conclude that the quantity $F_1P + F_2P$ is also independent of P. □

This proof is clearly the best way to think about the problem; it is interesting that it was only discovered in 1822. Since the bounded (elliptic) conic sections are related to planetary orbits, one might wonder if parabolas and hyperbolas also play a role in astronomy. When Isaac Newton developed the Calculus in the 1660s he proved that Kepler's Three Laws can all be derived from the following more basic law.

Newton's Law of Universal Gravitation (1686). Every point mass attracts every other point mass with a force that is inversely proportionatl to the square of the distance between them. //

The concept of "force" is not obviously geometric[17] so I will end the story here. In closing, I will just note the following fact. By using Calculus, Newton was able to show that the "inverse square law" leads inevitably to elliptic planetary orbits. The same analysis also shows that a massive body traveling near the Sun will follow a **hyperbolic path** if it has too much energy to get trapped in a closed orbit. If its energy level is right on the cusp between between being trapped and being free then it will follow a **parabolic path**. What can't geometry do?

[17] but c.f. Einstein's General Relativity



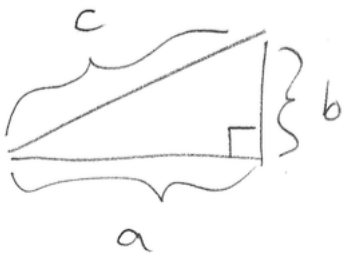
2 Euclidean and Non-Euclidean Geometry

2.1 The Deductive Method

In the last chapter I mentioned the birth of "capital M" Mathematics on the eastern shore of the Aegean sea in the 6th century BC. The founder of this school was apparently Thales of Miletus (c. 624–546 BC). For whatever reason, the Milesians began to question their world in a new way. Instead of just accepting traditional/mythological explanations, they developed the idea that the *kosmos* (the "ordered universe") **can be explained through careful thinking and reasoning**.

They adopted mathematical reasoning as the prototype for this new discipline. If a mathematical statement is *true* then they believed that its truth should be explained by systematic logical *proof*. We saw several proofs in the previous chapter. What did they have in common?

For example, let's consider the proof of the Pythagorean Theorem. Consider any right-angled triangle with side lengths a, b and c, as in the following figure:



In this case the Pythagorean Theorem says that $a^2 + b^2 = c^2$. I gave a logical argument for this fact and I succeeded in convincing most of you that it is **true**. More precisely, I succeeded in convincing you that the following implication is true:

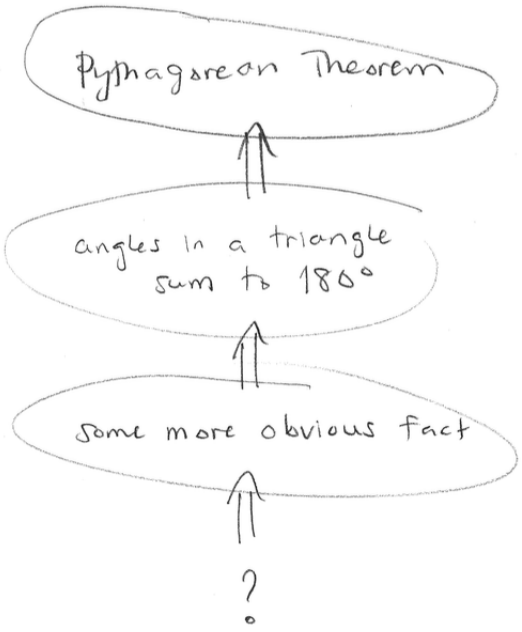
if the angles in any triangle sum to 180°, **then** the Pythagorean Theorem is true.

Since most of you **believe** that the angles in any triangle sum to 180° you were satisfied with this and so I ended the proof there. But the Milesian spirit says that we should also question **why** the angles in any triangle sum to 180°.

In modern mathematics we often express logical implication with a bold arrow as follows:

"$P \Rightarrow Q$" means " if P is true then Q is true ".

Thus I can express my proof in the following symbolic form:



If you do not already believe that the angles in any triangle sum to 180° then I will have to prove this to you by showing that it follows logically from some more obvious fact.

When does a proof end? In practice, the proof will continue until my intended audience is convinced. In principle, the proof will continue until the desired result is shown to follow from some basic facts that are "self-evidently true", i.e., facts that do not need to be proved. Thus the possibility of mathematical proof depends on the existence of some collection of "self-evident truths" that can be used as a foundation.

Terminology. In modern mathematics we use the word *axiom* for a true statement that does not need to be proved. A *deductive system* consists of:

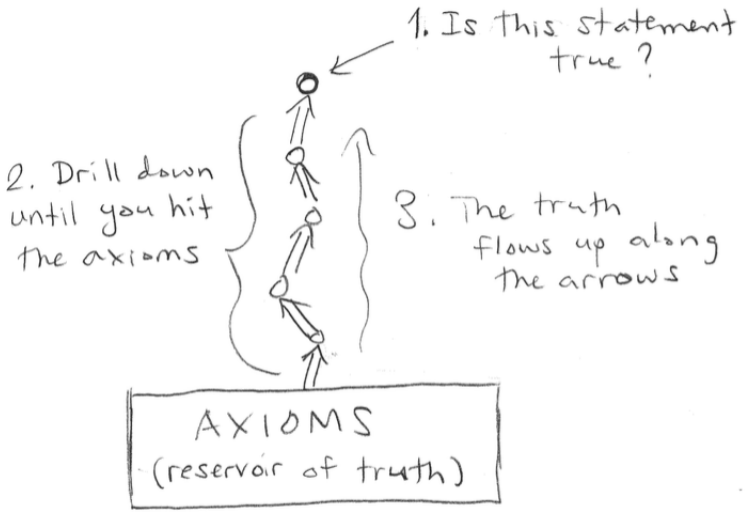
- a collection of axioms,
- some rules of logic telling us how we are allowed to combine them.

In principle the axioms can be chosen arbitrarily but in practice we want our deductive system to have the following properties:

- the axioms are intuitively true,
- the axioms are simple and easy to use,
- there **enough axioms** to prove all of the results we want (the axioms are *complete*),
- but there are **not too many axioms** (the axioms are *independent*, i.e., no one axiom can be proved from the others).

If a mathematical statement follows logically from the axioms, then we say that it is *true* and we call it a *theorem*. //

I visualize a deductive system with the following schematic diagram:



Here the arrows "⇒" represent "logical implication". To prove a given mathematical statement, the goal is to "drill downwards" by constructing a chain of logical implications from progessively more basic statements. If we can reach the axioms, then "truth" will **flow upward** along the chain of implications, showing that the original statement is true.

The Greek word *mathema* means "that which is learnt" and the Pythagoreans used the word *mathematikoi* to refer to "teachers". The community of mathematikoi grew dramatically in the years 600–300 BC. At first their choice of axioms would have been ad hoc, and they would have used differing standards of mathematical rigor. Eventually, however, as the community spread across the Greek world they would have felt a need to standardize their methods. After all, the main achievement of a mathematical proof is to compel different individuals (perhaps even enemies) to accept the truth of a given statement; if they can't agree on the underlying standards then the proof is less compelling.

Finally, around 300 BC, a systematic treatise was written bringing all of Greek mathematics under a single deductive system. This treatise was so well written that it became the standard mathematical textbook in the Western world, and its logical methods were so compelling that mathematicians accepted them unquestioningly until well into the 19th century.

2.2 Euclid's *Elements*

In the late Classical world, the center of intellectual activity moved from Athens, Greece, to Alexandria, Egypt. Alexander the Great (c. 356–323 BC) founded the city in c. 331 BC and then after his death it was ruled as a kingdom by his general Ptolemy I (c. 367–283 BC). This Ptolemy[18] was a patron of letters who ordered the construction of the Museum of Alexandria, which contained the great Library of Alexandria.

The Museum functioned as a research institute and it attracted scholars from all over the Hellenistic world. One of these scholars was the mathematician Euclid of Alexandria (fl. 300 BC) who wrote the definitive treatise on Greek mathematics, called the *Elements*. We know almost nothing about Euclid the man[19] but his monumental work became the most important mathematical treatise in history.

Euclid's *Elements* consists of 13 books and contains a total of 468 theorems (called *propositions*). Each book covers a slightly different topic, spanning the whole range of Greek mathematics. Here are a summaries of the first and the final book:

> Book I (with 48 propositions) is an exhaustive proof of the **Pythagorean Theorem**. The theorem itself is the subject of propositions I.47 and I.48; the first 46 propositions slowly build up to the Pythagorean Theorem starting from first principles. For example, Prop I.32 says that the interior angles of any triangle sum to 180°.

[18] not to be confused with the astronomical Ptolemy who lived 400 years later

[19] The 20th century collective of French mathematicians who wrote under the name "Bourbaki" suggested that "Euclid" might have been a pseudonym for a similar collective of mathematicians.

Book XIII (with 18 propositions) is devoted to the construction and classification of the five **Platonic solids**. The hardest of these to construct is the regular dodedecahedron; this is done in Proposition XIII.17. Then Proposition XIII.18 compares the five Platonic solids proves that no other regular polyhedra are possible.

Each book also contains *definitions* for the concepts it will use. Here is a selection of definitions from Book I:

Definition I.1. A *point* is that which has no part.

Definition I.2. A *line* is breadthless length.

Definition I.4. A *straight line* is a line which lies evenly with the points on itself.

Definition I.15. A *circle* is a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure equal one another.

Definition I.16. And the point is called the *center* of the circle.

Some of the definitions are clearly important (such as the definition of a circle) but some of them seem unnecessary (such as the definition of a line).

All 13 volumes form a single deductive system based on a collection of 10 axioms, which are listed at the beginning of Book I. The axioms are divided into two kinds: 5 "postulates" and 5 "common notions". In the previous section we discussed how important it is to select good axioms; so you might be curious which axioms Euclid chose as a foundation for all of Greek mathematics.

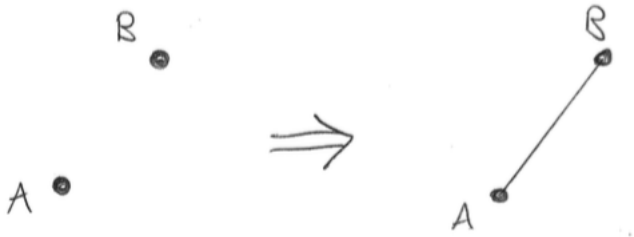
Here we should recall the "crisis of incommensurables" that occurred when Pythagorean mathematicians discovered the existence of "irrational numbers". This may be the reason that Euclid chose to found his system on **straight lines and circles** instead of on **numbers**.[20]

Euclid's Postulates. All of the propositions in Euclid are stated in terms of geometric constructions. The postulates tell us exactly what kind of geometric constructions are allowed; they are based on an idealized "straightedge and compass".

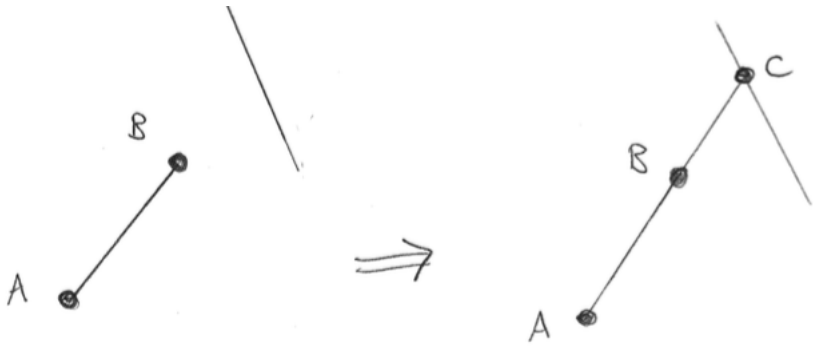
Let the following be postulated:

Postulate 1. *To draw a straight line from any point to any point.* This is the function of the straightedge:

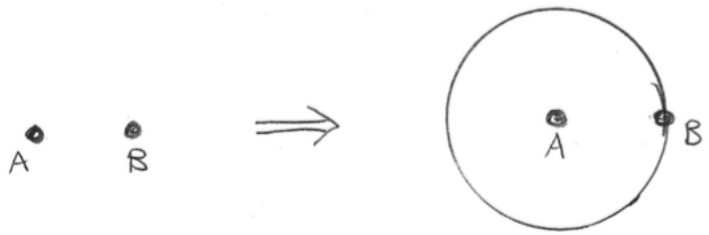
[20]It is also possible that the "Alexandrian numeral system" (a precursor of Roman numerals) was too awkward to allow for a systematic development of algebra. Our modern decimal place system comes from India. It migrated through the Arabic world and eventually came to Europe around 1200 AD.



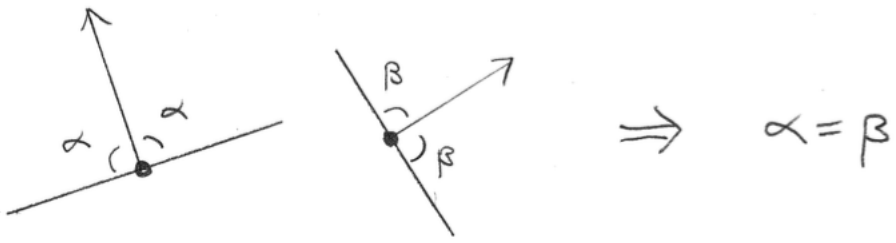
Postulate 2. *To produce a finite straight line continuously in a straight line.* Greek mathematicians avoided the concept of "completed infinity" in their mathematics and dealt only with "potential infinity". Thus every line in Euclid is only a finite line segment, but it can be extended when desired:



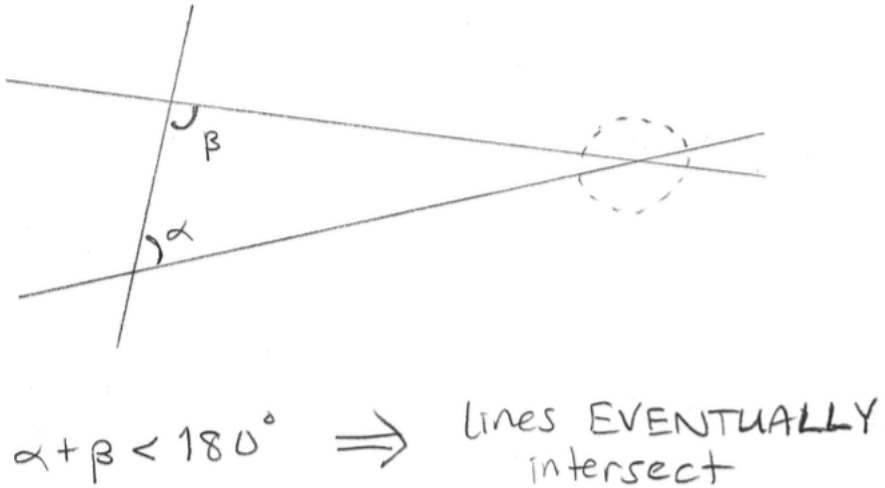
Postulate 3. *To describe a circle with any center and radius.* This is the function of the compass:



Postulate 4. *That all right angles equal one another.* Euclid had **defined** a right angle (in Definition 10) as half of a straight line angle. His goal is to use right angles as as the unit of angle measurement; this axiom allows him to compare angles at different points in the plane:



Postulate 5. *That, if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.* The 5th postulate became known as the "Parallel Postulate". We can visualize it as follows:



Euclid's Common Notions. These are the principles of **comparison** for geometric figures.

Common Notion 1. *Things which equal the same thing also equal one another.* In modern symbolic terms we would phrase this as follows:

$$\text{if } a = c \text{ and } b = c, \text{ then } a = b.$$

Common Notion 2. *If equals are added to equals, then the wholes are equal.* In modern terms:

$$\text{if } a = b \text{ and } c = d \text{ then } a + c = b + d.$$

Common Notion 3. *If equals are subtracted from equals, then the remainders are equal.* In modern terms:

$$\text{if } a = b \text{ and } c = d \text{ then } a - c = b - d.$$

Common Notion 4. *Things which coincide with one another equal one another.* This seems a bit mysterious to modern readers. I guess it means that if you move one geometric figure over another and they fit perfectly, then they must have equal lengths, angles, areas, volumes, etc.

Common Notion 5. *The whole is greater than the part.* In modern terms we would say that

$$0 < 1.$$

This is the foundation for comparison of magnitudes.

With just these 10 axioms, Euclid was able to rigorously re-derive most of the results of Greek mathematics, starting with the construction of an equilateral triangle in Prop I.1 and ending with the classification of the five Platonic solids in Prop XIII.18. The only major omission was the theory of *conic sections*. Apparently Euclid wrote a separate work on *Conics*. This work is lost but it served as the foundation for the *Conics* of Apollonius of Perga, which survives.

Before diving into the contents of Book I, let me briefly discuss what came **after** the *Elements* when it was eventually abandoned in the 19th century.

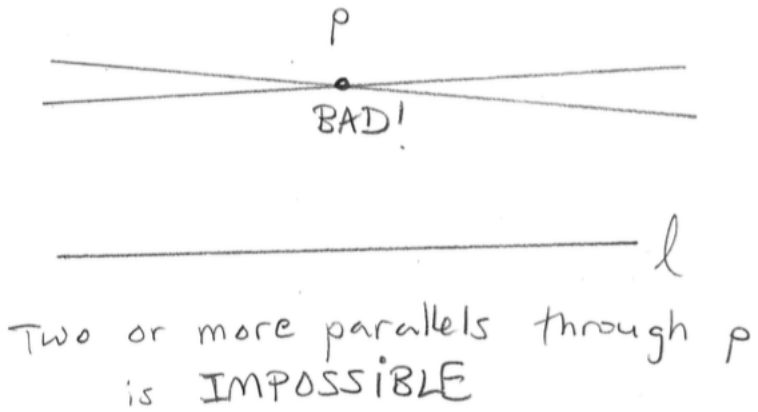
The 5th postulate (The Parallel Postulate) was always controversial. Some felt that it was not obvious enough to be an axiom.[21] For this reason many people tried to **prove** the Parallel Postulate from the other 9 axioms, i.e., to show that it is a **theorem** and not an **axiom**. This would fix the problem because theorems don't need to be obvious. But all attempts to prove the 5th postulate failed. All people were able to do was to replace it with logically-equivalent alternatives. The most famous reformulation is based on Euclid's Prop I.31:

Proposition I.31 [It is possible] *to draw a straight line through a given point parallel to a given straight line.* That is, given a line ℓ and a point p not on ℓ, it is possible to construct a line that contains p and never intersects ℓ.

In his commentary to the *Elements*, the Neo-Platonist philosopher Proclus Lycaeus (412–485 AD) proposed the following replacement. It was made famous by inclusion in John Playfair's textbook *Elements of Geometry* (1795):

Postulate 5' (Playfair's Postulate). *In a plane, given a line and a point not on it, at most one line parallel to the given line can be drawn through the point.* The existence of the line follows from the other 9 axioms; the key statement here is that there exists **at most one** such line:

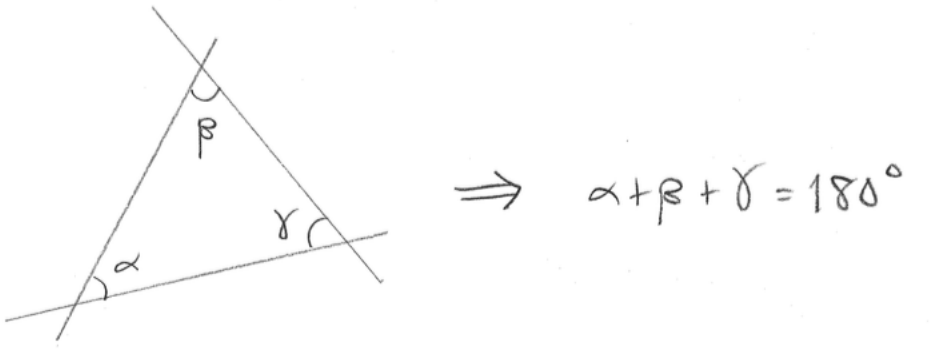
[21] What do *you* think?



By *logically-equivalent* I mean that Postulate 5' plus the other 9 axioms implies Postulate 5; and, similarly, Postulate 5 plus the other 9 axioms implies Postulate 5'. Thus we can choose the version that we like without affecting the truth of any other statement of Euclidean geometry.

It turns out that Postulates 5 and 5' are also equivalent to the following statement, which appears in Euclid as Proposition I.32:

Postulate 5" (The Triangle Postulate). *The sum of the angles of a triangle is two right angles.* Picture:



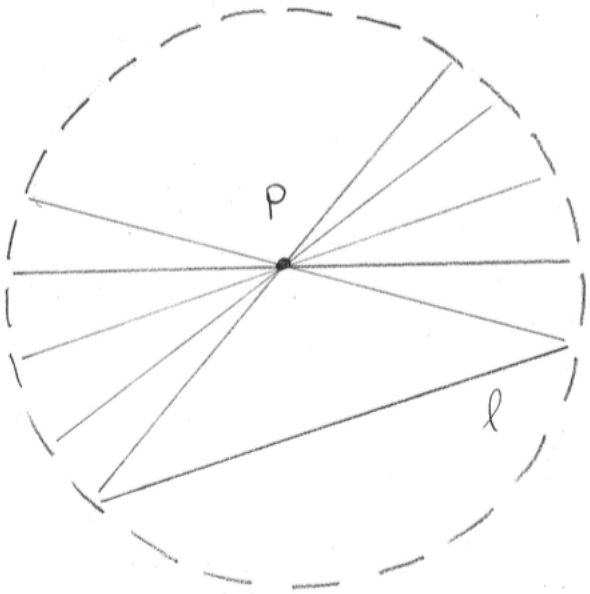
You might feel that this Postulate 5" is obviously true but I would disagree. The common problem with Postulates 5, 5' and 5" is that they all make a connection between things that could be very far apart in the plane. (For example, the triangle could be astronomically huge. In this case how could we ever measure and compare its angles?) The notion of "parallelism" itself is non-local because it states that two lines will **never intersect** no matter how far you extend them. This seems to go against the spirit of Greek mathematics, which tried to avoid the notion of "completed infinity".

The mathematician Carl Friedrich Gauss (1777–1855) was the first person to take seriously the idea that the 5th postulate **might actually be false**. For some time he was employed as a land surveryor and legend says that he measured the angles of a triangle between three mountain peaks to compute the sum of its angles.[22] However, he feared the "howls of the Boeotians" for questioning the dogma[23] of Euclidean geometry and so he kept these investigations to himself.

The 5th postulate was officially questioned by Nikolai Lobachevky (1830) and Janos Bolyai (1832), two relatively unknown mathematicians who had less to lose than Gauss in terms of professional reputation. Each of them claimed that Euclid's 9 axioms together with the **opposite of the Playfair Postulate** forms a consistent "non-Euclidean geometry". That is, they assumed that given a line ℓ and a point p not on ℓ, there exists **at least two** lines through p that are parallel to ℓ. The status of this non-Euclidean geometry was dubious until Eugenio Beltrami (1868) and Felix Klein (1873) proved the following result:

If Euclidean geometry is consistent (i.e., free from logical contradiction) then the Bolyai-Lobachevky geometry is also consistent.

The Beltrami-Klein Disk Model. To prove that non-Euclidean geometry is consistent, they found a way to model non-Euclidean geometry **inside of** Euclidean geometry. They did this by slightly redefining the notions of "point", "line" and "circle". Consider a fixed circle in the Euclidean plane:



[22] He must have found that the sum was within the margin of error of 180° otherwise we would have heard more about it.

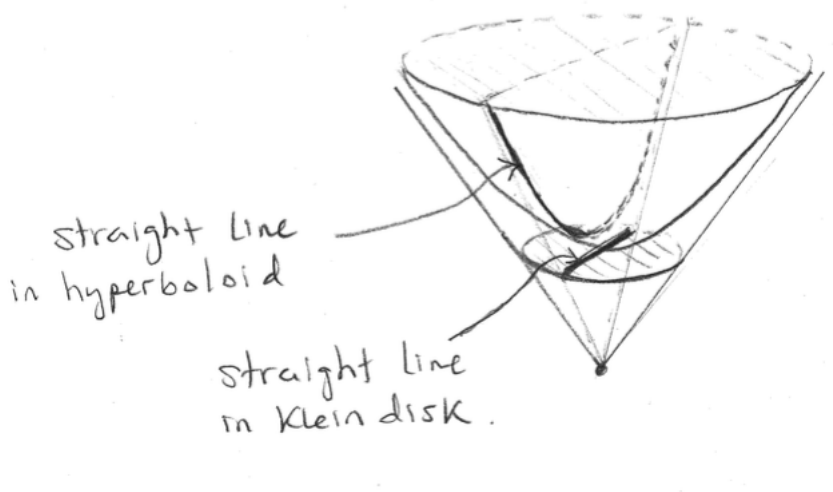
[23] The philosopher Immanuel Kant (1724–1804) had claimed that Euclidean geometry is a *synthetic a priori*, i.e., an absolute truth that is independent of human experience.

They redefined "point" to mean "a point in the interior of the disk" and they redefined "line" to mean "a line segment in the interior of the disk". In this way we can see that the Playfair Postulate is **false** because there exist **infinitely many** lines through a given point p that do not intersect a given line ℓ. In essence, the boundary of the disk (which is **not** included in the geometry) represents "infinity".

Their redefinition of "circle" is too difficult to state here (to make everything consistent they had to squash circles near the boundary), but suffice it to say that their new definitions of "point", "line" and "circle" satisfied all 9 of the other Euclidean axioms. It follows from this that the 5th postulate **can not be proved** and it **can not be disproved** from the other 9 axioms. Thus we are free to modify it as we please. //

One aspect of the Beltrami-Klein model that seems unnatural is the artificial boundary of the disk, which is treated as "infinity". Klein showed that the disk model is just the shadow of a more natural model with no boundary. Because of this model Klein invented the term *hyperbolic geometry* to refer to all of the various non-Euclidean models.

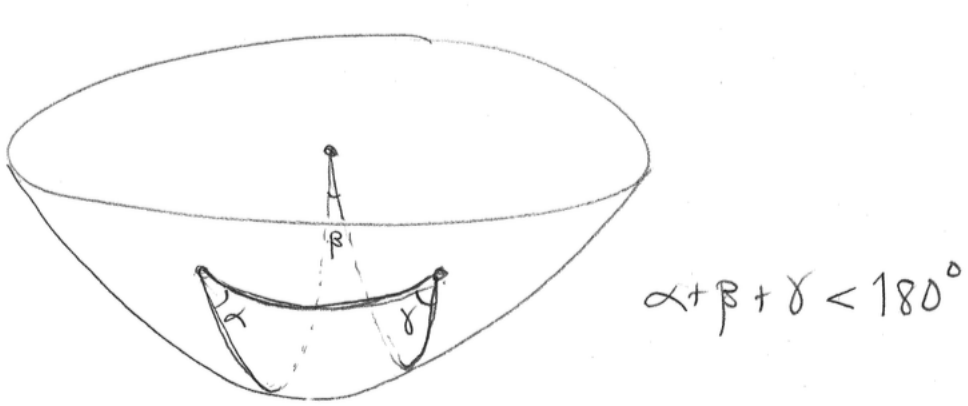
The Hyperboloid Model. If we rotate a hyperbola around its axis of symmetry then we obtain a bowl-shaped surface called a *hyperboloid* sitting on a circular disk inside of a circular cone:



The "points" of the geometry are the "points on the hyperboloid" and the "straight lines" are "intersections of the hyperboloid with a plane through the apex". The disk sitting under the hyperboloid is the Klein disk, whose straight lines are the intersections with the same planes.

The nice thing about this model is that it preserves the geometric properties of the Klein disk but it gets rid of the artificial boundary circle. The hyperboloid is a truly infinite surface with

a non-Euclidean geometry.[24] By putting three "straight lines" together we can also see[25] that *hyperbolic triangles* have angles that sum to **less than** 180°:



And this is certainly expected because we know that the Triangle Postulate is logically-equivalent to Playfair's Postulate, which is false in this model. //

The hyperbolic model provides the following insight:

Non-Euclidean geometry is caused by "curvature".

Euclidean geometry implicitly assumes that the universe is "flat"; and this does seem to be true on small scales. But hyperbolic geometry shows us that a locally-flat space might have large-scale curvature. For example, a small two-dimensional person living on a hyperboloid would not be able to distinguish their universe from a flat plane. Similarly, our three-dimensional universe seems to be locally-Euclidean. However, it is at least logically possible that the universe is curved at large scales.

Einstein's *General Relativity* is based on a very similar idea: that gravity in our three-dimensional space is caused by curvature in four-dimensional *space-time*. The large-scale geometry of the purely spatial universe, i.e., whether it has global curvature or not, is still an open question. In any case, the universe is close enough to Euclidean on small scales that Gauss couldn't tell the difference.

2.3 Selections from Book I

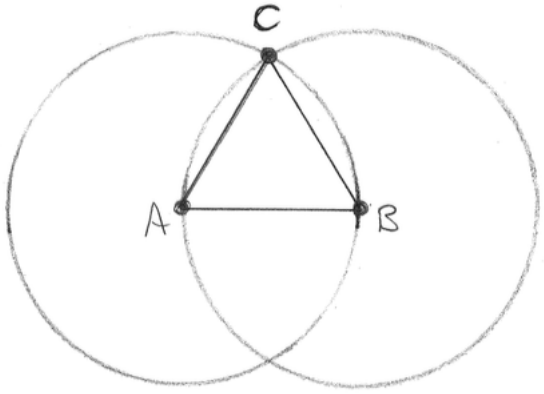
Now let's take a look at some specific theorems from Book I of Euclid's *Elements*.

Proposition I.1. *To construct an equilateral triangle on a given finite straight line.*

[24] Unfortunately we still have to modify the notions of "distance" and "circle". It follows from theorems of Gauss and Hilbert that there is no *distance-preserving* Euclidean model of the non-Euclidean plane.

[25] I'm lying a bit here because the "angles" also get distorted.

Proof. Let AB be the given straight line.



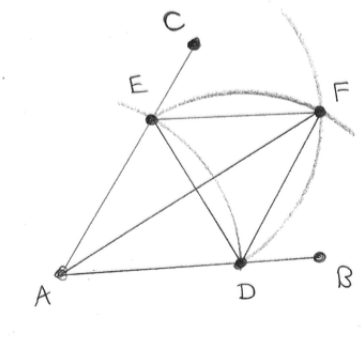
It is required to construct an equilateral triangle on the straight line AB. To do this we first construct the circle with center A and radius AB [Postulate 3] and the circle with center B and radius AB [Postulate 3]. Let C be one point of intersection of the two circles. [Euclid doesn't say why the intersection exists; he just assumes that the picture isn't lying.] Now construct the triangle ABC [Postulate 1].

Now I claim that ABC is an equilateral triangle. From the definition of "circle" [Definition 15] we must have $AB = AC$ since these are two radii of the same circle. Similarly, we must have $AB = BC$. Therefore each of the lengths AC and BC equals the length AC. Finally, Common Notion 1 implies that $AC = BC$. Q.E.D.

On HW1 I have asked you to come up with similar constructions for regular squares, hexagons and octagons. The following proposition will be helpful for these constructions.

Proposition I.9. *To bisect a given rectilinear angle.*

Proof. Let $\angle BAC$ be the given rectilinear angle.



It is required to bisect it.

Take an arbitrary point D on AB. Cut off AE from AC equal to AD. [Euclid quotes Proposition I.3 to do this, but you can just as easily draw the circle with center A and radius AD and let E be its point of intersection with AC.] Now construct the equilateral triangle $\triangle DEF$ using Proposition I.1. I say that the angle $\angle BAC$ is bisected by the straight line AF.

To show this I will first show that the triangles $\triangle ADF$ and $\triangle AEF$ are congruent. Indeed, we know that $AD = AD$ by contruction, $AF = AF$ by coincidence and $DF = EF$ by Proposition I.1. Since the two triangles have equal side lengths they are congruent by Proposition I.8. [This is the side-side-side criterion for congruence. We omit the proof.] Finally, since the two triangles are congruent we conclude that

$$\angle EAF = \angle DAF$$

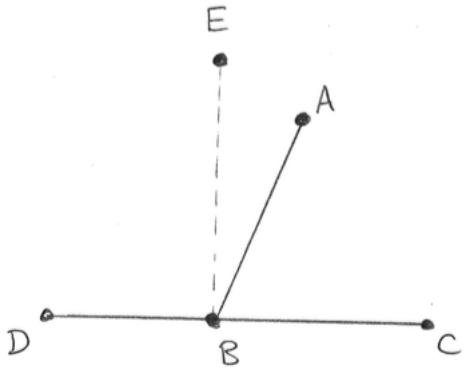
as claimed. Q.E.D.

For the rest of the section we will follow Euclid's proof that the interior angles of any triangle sum to 180°.[26] Of course we know that this argument must involve the Parallel Postulate in some way; it will make its first appearance in the proof of Proposition I.29 below.

Euclid's proof is long-winded so I will just show the key steps and I will paraphrase some of the proofs in modern language. We begin with Proposition I.13.

Proposition I.13. *If a straight line stands on a straight line, then it makes either two right angles or angles whose sum equals two right angles.*

Euclid's Proof. Let AB be a straight line standing on the straight line CD and consider the two angles $\angle ABC$ and $\angle ABD$.



[26] Instead of 180°, Euclid would say "two right angles". The convention of dividing the circle into 360° comes from the ancient Babylonians, who used a base 60 numeral system.

I claim that $\angle CBA + \angle ABD$ equals two right angles.

If $\angle CBA = \angle ABD$ then each of them is a right angle [Definition 10] so we are done. Otherwise we can assume without loss of generality that $\angle ABD$ is greater than a right angle. Draw the line BE at right angles to CD [which is possible by Proposition I.11, omitted]. Since $\angle CBE = \angle CBA + \angle ABE$ and $\angle CBE = \angle EBD$, Common Notion 2 tells us that

$$\angle CBE + \angle EBD = \angle CBA + \angle ABE + \angle EBD.$$

And then since $\angle DBA = \angle DBE + \angle EBA$, Common Notion 2 tells us that

$$\angle DBA + \angle ABC = \angle DBE + \angle EBA + \angle ABC.$$

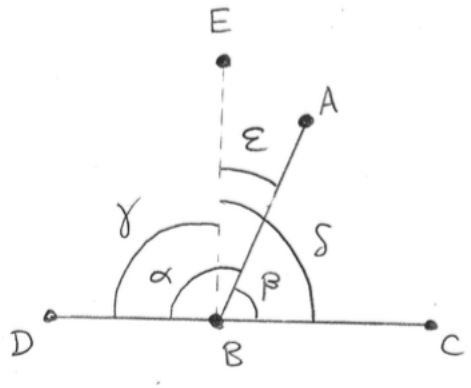
Finally, since $\angle DBA + \angle ABC$ and $\angle CBE + \angle EBD$ are both equal to the same thing (shown in the previous two equations), Common Notion 1 tells us that

$$\angle DBA + \angle ABC = \angle CBE + \angle EBD,$$

which is just what we wanted to show. Q.E.D.

That was Euclid's proof.[27] Here's my cleaned up version.

My Proof. Label the angles in Euclid's diagram as follows:



We want to show that $\alpha + \beta = 180^\circ$.

By assumption we know that $\gamma = \delta = 90^\circ$ and hence $\gamma + \delta = 180^\circ$. But then we have

$$\begin{aligned} \alpha + \beta &= (\gamma + \varepsilon) + \beta \\ &= \gamma + (\varepsilon + \beta) \\ &= \gamma + \delta \end{aligned}$$

[27] Except that I used the modern symbols "+" and "=" where Euclid said "the sum of" and "is equal to".

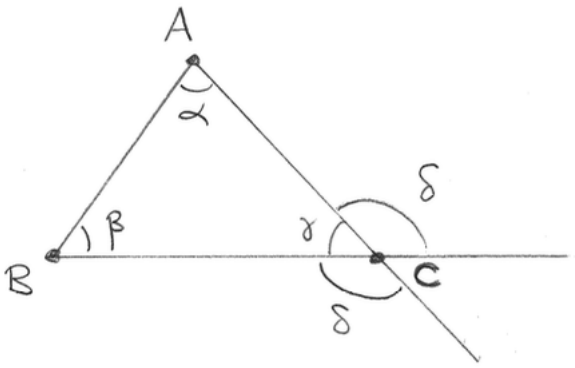
$$= 180^\circ$$

as desired. □

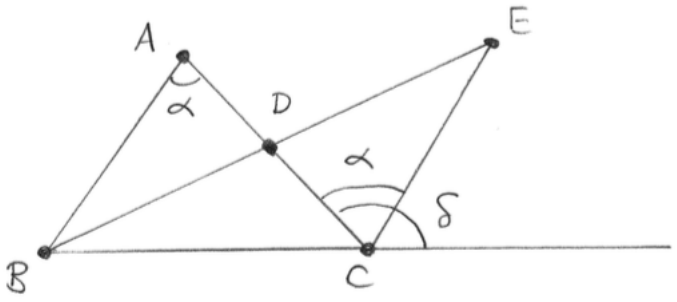
I think my version is easier to read. Euclid didn't have a symbolic notation for algebra so he had to express the sequence of equations verbally.

Proposition I.16. *In any triangle, if one of the sides is produced, then the exterior angle is greater than either of the interior and opposite angles.*

Proof. Consider a triangle $\triangle ABC$ with interior angles α, β, γ and extend the lines AC and BC to produce the exterior angle δ, as in the following figure:



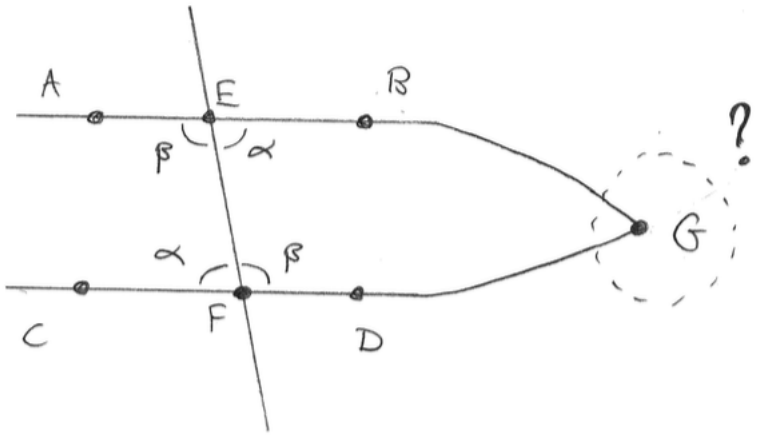
I claim that $\alpha < \delta$ and $\beta < \delta$. Since the proofs are similar I will just show that $\alpha < \delta$. To do this, let D be the midpoint of AC [Prop I.10] and extend the line AD to the point E such that $BD = DE$ [Prop I.3], as in the following figure:



Since $AD = CD$, $BD = DE$ and $\angle ADE = \angle CDE$ [from Prop I.15] we conclude that the triangles $\triangle ADB$ and $\triangle CDE$ are congruent. [This is the side-angle-side criterion for congruence from Prop I.4.] In particular we conclude that $\angle DCE = \angle DAB = \alpha$. Finally, since the angle $\alpha = \angle DCE$ is contained inside the exterior angle δ we conclude from Common Notion 5 [the whole is greater than the part] that $\alpha < \delta$. □

Proposition I.27. *If a straight line falling on two straight lines makes the alternate angles equal to one another, then the straight lines are parallel to one another.*

Proof. Suppose that the line EF crosses the two lines AB and CD, and assume that the alternate angles are equal, as in the following figure:

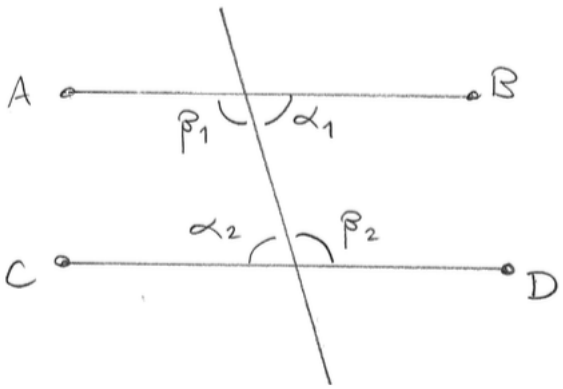


I claim that the lines AB and CD are parallel. To prove this, assume for contradiction that the lines are **not** parallel so they meet on one side. If the lines at G as in the figure then the interior angle $\beta = \angle EFG$ of the triangle $\triangle EFG$ is equal to the exterior angle $\beta = \angle AEF$, which contradicts Proposition I.16 from above. If the lines meet on the other side we get a similar contradiction where α is both an interior and an exterior angle for some triangle. □

So far all of these results are also true in hyperbolic geometry. The next result (Prop I.29) will be our first result that uses the Parallel Postulate.

Proposition I.29. *A straight line falling on parallel straight lines makes the alternate angles equal to one another, the exterior angle equal to the interior and opposite angle, and the sum of the interior angles on the same side equal to two right angles.*

Proof. Let AB and CD be parallel lines and let them be cut by a straight line making alternate angles $\alpha_1, \alpha_2, \beta_1, \beta_2$ as in the following figure:



In this case I claim that $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$.

To prove this, assume for contradiction that $\alpha_1 \neq \alpha_2$. Without loss of generality we will suppose that $\alpha_1 < \alpha_2$. Then since $\alpha_1 + \beta_1 = 180^\circ$ and $\alpha_2 + \beta_2 = 180^\circ$ [by Prop I.13 above] we conclude that

$$\begin{aligned} \alpha_1 &< \alpha_2 \\ \alpha_1 + \beta_2 &< \alpha_2 + \beta_2 \\ \alpha_1 + \beta_2 &< 180^\circ. \end{aligned}$$

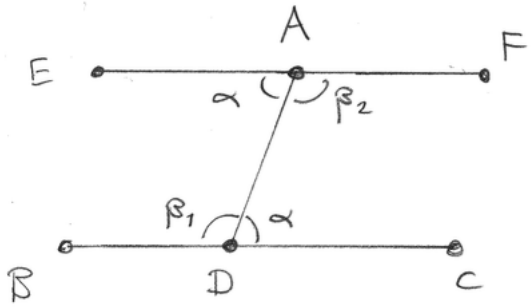
But then the Parallel Postulate [Postulate 5] says that that AB and CD are not parallel, which is the desired contradiction. We conclude that $\alpha_1 = \alpha_2$ and then it follows from Proposition I.13 that

$$\beta_1 = 180^\circ - \alpha_1 = 180^\circ - \alpha_2 = \beta_2.$$

□

Proposition I.31. *To draw a straight line through a given point parallel to a given straight line.*

Proof. Let A be the given point and let BC be the given line:

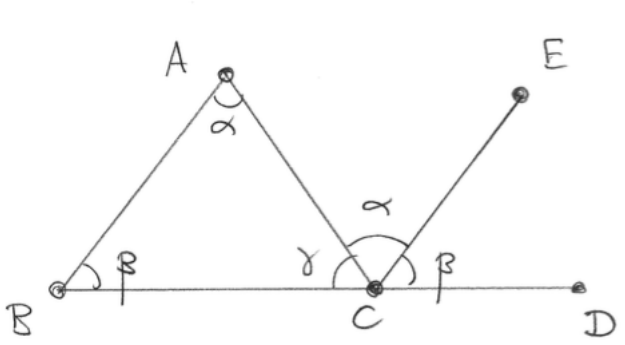


Now choose a random point D on BC and draw the segment AD. By Proposition I.23 [omitted] it is possible to construct a segment AE so that angle $\alpha = \angle EAD$ equals angle $\alpha = \angle CDA$. Now extend the segment EA to F and define the angle $\beta_2 = \angle DAF$. Then Proposition I.13 implies that $\beta_1 = \beta_2$ as in our proof of Proposition I.29 above. Finally, since the alternate angles are equal we conclude from Proposition I.27 above that the line EF is parallel to BC as desired. □.

We have arrived at the big theorem of this section.

Proposition I.32 (The Angles in a Euclidean Triangle Sum to 180°). *In any triangle, if one of the sides is produced, then the exterior angle equals the sum of the two interior and opposite angles, and the sum of the three interior angles of the triangle equals two right angles.*

Proof. Consider a triangle $\triangle ABC$ with internal angles α, β, γ as in the following figure:



I claim that $\alpha + \beta + \gamma = 180^\circ$.

To prove this we extend the line BC to an arbitrary point D and we use Proposition I.31 (which does not need the Parallel Postulate) to draw a line segment CE parallel to AB. Now we apply Proposition I.29 (which **does** need the Parallel Postulate) two times:

(1) When the line AC crosses the parallel lines AB and CE it makes the alternate angles $\alpha = \angle BAC$ and $\angle ECA$ equal.

(2) When the line BD crosses the parallel lines AB and CE it makes the alternate angles $\beta = \angle ABC$ and $\angle ECD$ equal.

Finally, since the angles $\gamma = \angle BCA$, $\alpha = \angle ACE$ and $\beta = \angle ECD$ add up to the straight line BD we conclude that

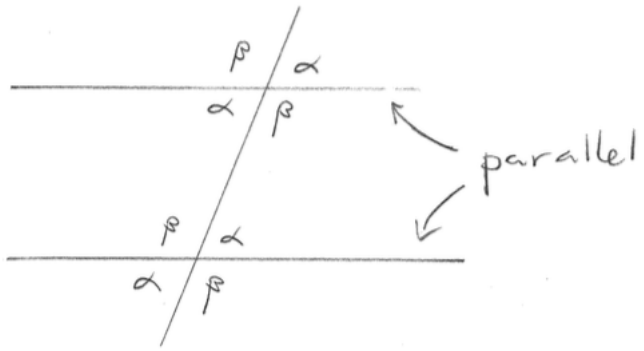
$$\alpha + \beta + \gamma = 180^\circ$$

as desired. □

This finally completes the proof of the Pythagorean Theorem that I presented in in Chapter 1.1. The rest of Euclid's Book I is devoted to his own proof of the Pythagorean Theorem, which I find less convincing than our proof so I won't discuss it here.

2.4 Triangles and Curvature

In the previous section we saw Euclid's proof that the angles in a triangle sum to 180°. This was his Proposition I.32. We also saw that Euclid's proof depends on the Parallel Postulate. In particular, he needed the Parallel Postulate to prove Prop I.29 which says that a straight line falling on two parallel lines makes alternate angles equal, as in the following figure:

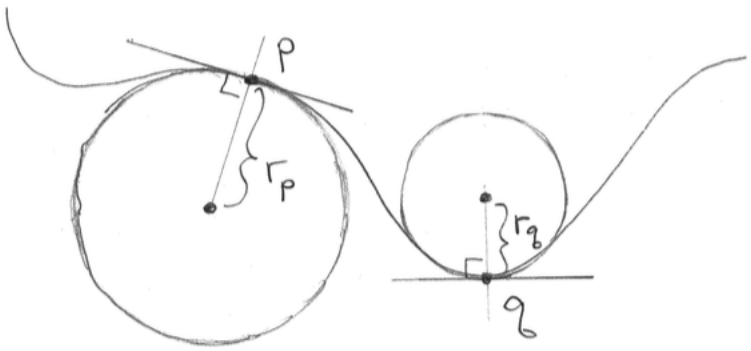


I also mentioned above that there exist consistent geometries in which the Parallel Postulate is **false** but the other 9 Euclidean axioms are **true**. This was first suggested by Bolyai and Lobachevsky in the early 1800s and then it was proved rigorously by Beltrami and Klein in the late 1800s. Klein called the new geometry *hyperbolic* because of a certain model based on the hyperboloid surface.

It is difficult to visualize hyperbolic geometry because of a famous theorem of Carl Friedrich Gauss (1828) which he called the *Theorema Egregium* (Latin for "remarkable theorem"). To discuss this theorem I have to show you Gauss' notion of *curvature* for paths and surfaces.

Definition of "Gaussian" Curvature. Consider a smooth path in space. For each point p on the path there is a unique circle (called the *osculating* circle) that fits the path near p better[28] than any other circle:

[28] We would need Calculus to make this precise.



If r_p is the radius of the osculating circle at p then Gauss defined the *curvature* of the path at p as the reciprocal of the radius:

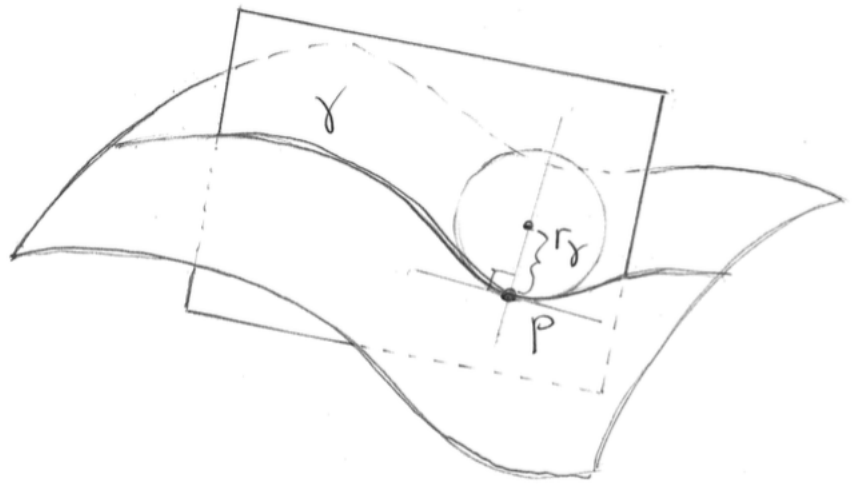
$$\kappa_p := 1/r_p$$

$$(\text{curvature of the path at } p) = \frac{1}{(\text{radius of the osculating circle at } p)}$$

If the path near p is a straight line segment then we say that the osculating circle at p has infinite radius ($r_p =$ "∞") and we define the curvature as $\kappa_p := 0 =$ "$1/\infty$". Thus at every point p on a smooth curve there is a non-negative curvature $\kappa_p \geqslant 0$. A straight line is a path of "constant curvature 0" and a circle of radius r is a path of "constant curvature $1/r$". Every other path has curvature that changes from point to point. For example, in the above figure we have $r_p > r_q > 0$ and hence $0 < \kappa_p < \kappa_q$. A point of "infinite curvature" would be very sharp; but this will never happen because we assume that the path is "smooth".

Now consider a point p on a smooth two-dimensional surface. There are infinitely many smooth paths in the surface that go through p and each of them has an osculating circle at p. Somehow we want to pull all of this information together to obtain a single number κ_p that represents the "curvature of the surface at the point p". Here is Gauss' idea:

Call one side of the surface "above" and the other "below" (at the end we will see that the choice is arbitrary). Now for each flat plane containing p, consider the *sectional path* γ where this plane intersects the surface, as in the following figure:



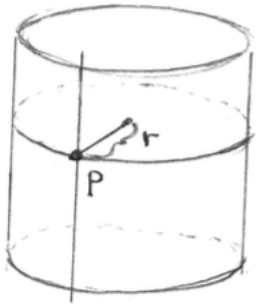
This path has an osculating circle of radius r_γ at p (possibly with $r_\gamma =$ "∞"). If the osculating circle is "above" the surface when we define the *sectional curvature* by $\kappa_\gamma = 1/r_\gamma$ and if the osculating circle is "below" the surface we define $\kappa_\gamma = -1/r_\gamma$. Thus we have a whole set of "sectional curvatures" at p:

$$\{\kappa_\gamma : \text{where } \gamma \text{ is a sectional path in the surface going through } p\}$$

This set might contain infinitely many different numbers or it might contain just one number (when the surface has "constant sectional curvature"). In any case, since the surface is smooth, Gauss observed that there is a **minimum** and a **maximum** sectional curvature satisfying $\kappa_1 \leqslant \kappa_2$. Then he defined the *(Gaussian) curvature* of the surface at p to be the product of these extreme values:

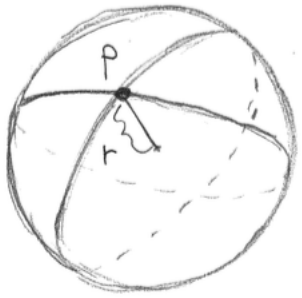
$$\kappa_p := \kappa_1 \cdot \kappa_2$$
$$\text{(curvature of the surface at } p) = (\text{min sectional curvature}) \cdot (\text{max sectional curvature})$$

If the surface is a flat plane near p then we have $\kappa_1 = \kappa_2 = 0$ and hence the Gaussian curvature at p is $\kappa_p = 0 \cdot 0 = 0$. Thus a flat plane is a surface of "constant sectional curvature 0". Now consider an infinite circular cylinder of radius r:



The two extreme sectional curvatures come from (1) sectional paths along the cylinder (which are straight lines) and (2) sectional paths perpendicular to the cylinder (which are circles of radius r). If the inside of the cylinder is "above" then the extreme curvatures are $\kappa_1 = 0 < 1/r = \kappa_2$ and the Gaussian curvature at any point p is $\kappa_p = 0 \cdot (1/r) = 0$. If the outside of the cylinder is "above" then the extreme curvatures are $\kappa_1 = -1/r < 0 = \kappa_2$ and the Gaussian curvature at any point is still $\kappa_p = (-1/r) \cdot 0 = 0$. In any case, we find that a cylinder is a surface of "constant Gaussian curvature 0" (even though the sectional curvatures are not all zero).

Next consider a point p on the surface of a sphere of radius r. In this case, every sectional path at p is a circle of radius r as in the following picture:



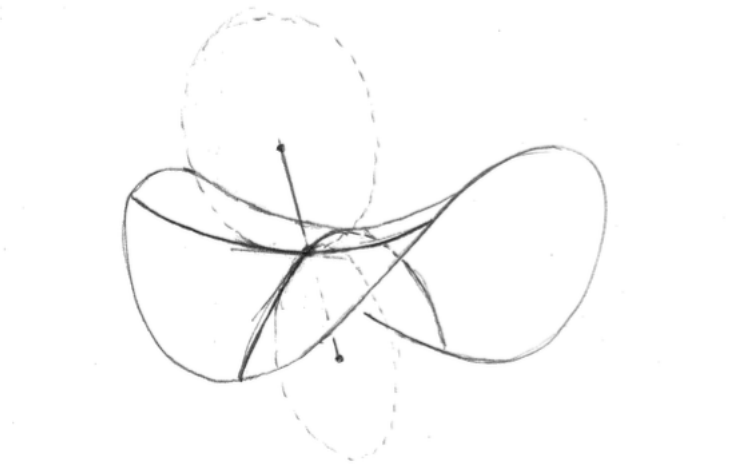
If we regard the inside of the sphere as "above" then all sectional curvatures are $1/r$; if the outside of the sphere is "above" then all sectional curvatures are $-1/r$. **In either case**, the Gaussian curvature at p is

$$\kappa_p = \frac{1}{r} \cdot \frac{1}{r} = \left(-\frac{1}{r}\right) \cdot \left(-\frac{1}{r}\right) = \frac{1}{r^2} > 0.$$

Thus a sphere of radius r is a surface of "constant positive Gaussian curvature $1/r^2$".

For the case of paths we observe that the curvature is always non-negative. However, for surfaces there is a notion of "negative curvature", which occurs when the two extreme osculating

circles are on opposite sides of the surface. For example, consider a point p on a saddle-shaped surface:



Regardless of which side of the saddle is "above", since the two extreme osculating circles are pointing in opposite directions we will find that the extreme sectional curvatures satisfy $\kappa_1 < 0 < \kappa_2$. Then it follows that the Gaussian curvature at p is a negative number: $\kappa_p = \kappa_1 \cdot \kappa_2 < 0$. Based on this example, any point of negative Gaussian curvature is called a *saddle point*. It is not clear whether there exists a surface of "constant negative Gaussian curvature". [We will return to this issue below.] //

Thus the Gaussian curvature of a surface is defined in terms of the osculating circles that **stick out of the surface at right angles**. If you were an ant living on the surface then it would be impossible to compute the curvature using this method. However, Gauss also proved a remarkable theorem showing that **the curvature can still be computed by ants on the surface**. To be specific, consider any two points p and q on a surface. A path of shortest length between p and q (that says within the surface) is called a *geodesic path*.[29] Then we define the *geodesic distance* by

$$d(p, q) := \text{length of a geodesic path between } p \text{ and } q$$

Gauss' *Theorema Egregium*[30] (1827). Let p be a point on a smooth two-dimensional surface. Then the Gaussian curvature κ_p can be computed if we know the gesodesic distances $d(p, q)$ for all of the points q near p. //

This is a subtle idea so we should think about it slowly. Let p be a point on a surface and consider any number $r \geqslant 0$. We define the "internal disk with center p and radius r" to be

[29] For example, the path that a plane must fly over the surface of the Earth. The Latin *geodaesia* literally means "Earth division".

[30] Latin for "remarkable theorem". What language do you think Gauss wrote in?

the collection of points q in the surface such that the geodesic distance satisfies $d(p,q) \leqslant r$. This is the region of the surface that can be reached by an ant if it starts at p and then walks for a distance r. Now let $A_p(r)$ be the **area** of this "internal disk".

If the surface is "flat" near p and if the radius r is small then we have the famous[31] formula

$$A_p(r) = \pi r^2.$$

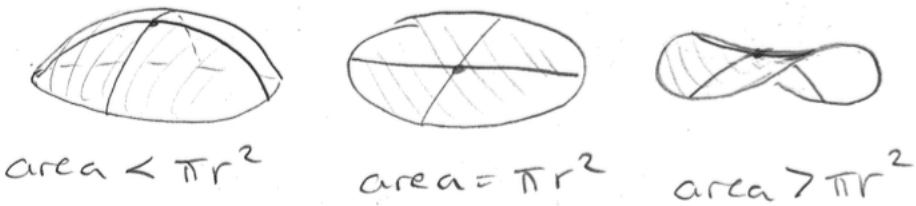
More generally, $A_p(r)$ is some function of the radius r that depends on the shape of the surface near p. This function can be computed from the geodesic distances $d(p,q)$ and a bit of Calculus. In the year 1848, three French mathematicians gave an elegant proof of Gauss' *Theorema Egregium* by establishing the following general formula for $A_p(r)$. For a point p on a smooth surface they showed that

$$A_p(r) = \pi r^2 - \kappa_p \cdot \frac{\pi r^4}{12} + \text{ higher order terms}, \tag{BDP}$$

where κ_p is the Gaussian curvature at p. The higher order terms are negligible when the radius r is small. This formula tells us that the curvature κ_p is determined by the area function $A_p(r)$. But the area function is determined by the geodesic distances $d(p,q)$; thus the curvature at p is also determined by the distances $d(p,q)$. Q.E.D.

There is a professor at Cornell named John Hubbard who likes to think of the Bertrand-Diquet-Puiseux formula (BDP) in terms of goats. Suppose we tie up our goat in a grassy meadow using a rope of length r. Thus the goat has an area $A_p(r)$ of grass to eat. In particular:

- If the meadow is flat ($\kappa_p = 0$) then the goat has πr^2 of grass to eat,
- If the meadow is in a valley or on a hilltop ($\kappa_p > 0$) then there is **less than** πr^2 to eat.
- If the meadow is in a mountain pass ($\kappa_p < 0$) then there is **more than** πr^2 to eat.



The *Theorema Egregium* also has an important consequence for map-making. The Earth is approximately a sphere. In any case, the surface of the Earth has **positive** Gaussian curvature. But a map, even if it is bent or folded as in the case of a cylinder, has **zero** Gaussian curvature. Since the curvature is a function of geodesic distances, Gauss' theorem

[31] I'm sure you already believe this formula. If not, I'll prove it to you in the next chapter.

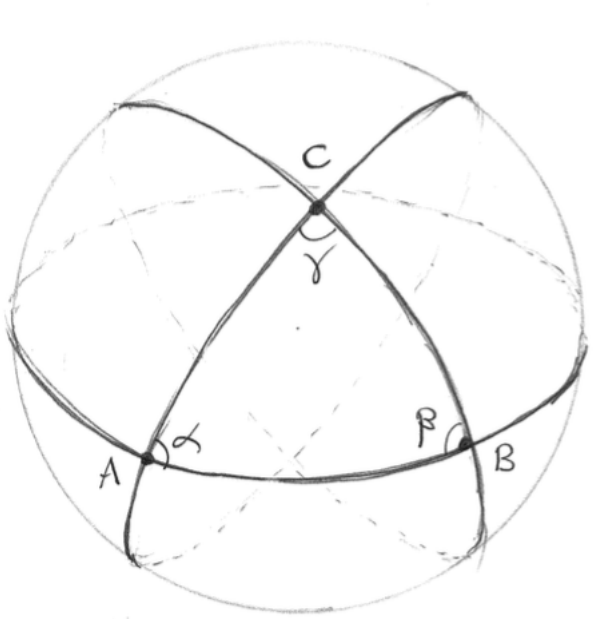
tells us that is it impossible to make a make a map of the Earth without distorting distances. In other words:

It is impossible to make a distance-preserving map of the Earth.

The problem is worse for larger regions of the globe, but even a relatively small region such as the state of Florida has enough curvature so that flat maps will be noticeably distorted. The history of map-making was a series of compromises based on the purpose that each map would be used for. Today the main problem is to stitch together flat satellite photographs into an accurate curved picture. //

But there is yet one more way to think about curvature which is intimately related to non-Euclidean geometry. This has to do with the sum of the angles in a geodesic triangle.

To begin we will consider a geodesic triangle on the surface of a sphere. Suppose that the vertices are A, B, C and the angles are α, β, γ as in the following picture:



If the sphere has radius r then each geodesic is a *great circle* of radius r. Observe that if we extend the sides of the triangle then they wrap around the back of the sphere to decompose the surface into 8 triangular regions. It we are clever than we can use this decomposition to compute the area of the triangle. The following theorem is attributed to the French mathematician Albert Girard (1629) and the English polymath[32] Thomas Harriot (1603).

Theorem (Area of a Spherical Triangle). Consider a geodesic triangle on the surface of

[32]Among other things, he traveled with Sir Walter Raleigh to the new world where they founded the colony of *Virginia* in honor of "the Virgin Queen" Elizabeth I.

a sphere of radius r. If the internal angles are α, β, γ then the angle sum $\alpha + \beta + \gamma$ is strictly greater than 180°. Furthermore, the area of the triangle is given by the following formula:

$$(\text{area of a triangle}) = (\text{angle excess}) \cdot r^2 = (\alpha + \beta + \gamma - \pi) \cdot r^2$$

Here we use the "radian measure"[33] π instead of 180° because we need to compare the angles to the radius of the sphere. //

Proof. Let $\triangle_{\alpha,\beta,\gamma}$ denote the area of the spherical triangle with internal angles α, β, γ, as shown in the figure above. Note from the figure that by extending the sides of the triangle we decompose the surface of the sphere into 8 triangles, and that these triangles come in opposite pairs of equal area. We will let $\triangle_\alpha, \triangle_\beta, \triangle_\gamma$ denote the areas of the triangles that occur across the sides of the original triangle and opposite the angles α, β, γ, respectively. Thus we must have

$$S = 2(\triangle_{\alpha,\beta,\gamma} + \triangle_\alpha + \triangle_\beta + \triangle_\gamma),$$

where S is the surface area of the whole sphere.

On the other hand, observe that the two triangles $\triangle_{\alpha,\beta,\gamma}$ and $\triangle_\alpha$ fit together to form a "lune" with angle α. By looking at the axis through the two endpoints of the lune we see that the lune covers a proportion $\alpha/2\pi$ of the full surface of the sphere. Thus we have

$$\triangle_{\alpha,\beta,\gamma} + \triangle_\alpha = \frac{\alpha}{2\pi} \cdot S.$$

Similarly, we have lunes with angles β and γ:

$$\triangle_{\alpha,\beta,\gamma} + \triangle_\beta = \frac{\beta}{2\pi} \cdot S \qquad \text{and} \qquad \triangle_{\alpha,\beta,\gamma} + \triangle_\gamma = \frac{\gamma}{2\pi} \cdot S.$$

Adding these three equations together gives

$$\begin{aligned}
3\triangle_{\alpha,\beta,\gamma} + \triangle_\alpha + \triangle_\beta + \triangle_\gamma &= \frac{(\alpha + \beta + \gamma)}{2\pi} \cdot S \\
6\triangle_{\alpha,\beta,\gamma} + 2\triangle_\alpha + 2\triangle_\beta + 2\triangle_\gamma &= \frac{(\alpha + \beta + \gamma)}{\pi} \cdot S \\
4\triangle_{\alpha,\beta,\gamma} + 2\triangle_{\alpha,\beta,\gamma} + 2\triangle_\alpha + 2\triangle_\beta + 2\triangle_\gamma &= \frac{(\alpha + \beta + \gamma)}{\pi} \cdot S \\
4\triangle_{\alpha,\beta,\gamma} + 2(\triangle_{\alpha,\beta,\gamma} + \triangle_\alpha + \triangle_\beta + \triangle_\gamma) &= \frac{(\alpha + \beta + \gamma)}{\pi} \cdot S,
\end{aligned}$$

and then substituting $S = 2(\triangle_{\alpha,\beta,\gamma} + \triangle_\alpha + \triangle_\beta + \triangle_\gamma)$ from the original equation gives

$$\begin{aligned}
4\triangle_{\alpha,\beta,\gamma} + 2(\triangle_{\alpha,\beta,\gamma} + \triangle_\alpha + \triangle_\beta + \triangle_\gamma) &= \frac{(\alpha + \beta + \gamma)}{\pi} \cdot S \\
4\triangle_{\alpha,\beta,\gamma} + S &= \frac{(\alpha + \beta + \gamma)}{\pi} \cdot S
\end{aligned}$$

[33] If this is not familiar to you, just wait; we'll discuss it soon.

$$
\begin{aligned}
4\triangle_{\alpha,\beta,\gamma} &= \frac{(\alpha+\beta+\gamma)}{\pi}\cdot S - S \\
4\triangle_{\alpha,\beta,\gamma} &= \frac{(\alpha+\beta+\gamma)}{\pi}\cdot S - \frac{\pi}{\pi}\cdot S \\
4\triangle_{\alpha,\beta,\gamma} &= \frac{(\alpha+\beta+\gamma-\pi)}{\pi}\cdot S \\
\triangle_{\alpha,\beta,\gamma} &= \frac{(\alpha+\beta+\gamma-\pi)}{4\pi}\cdot S.
\end{aligned}
$$

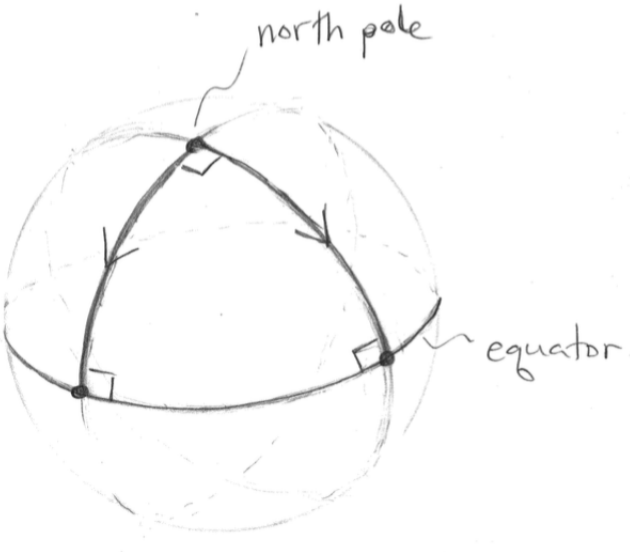
This is already enough to show us that the "angle excess" $(\alpha+\beta+\gamma-\pi)$ is **positive**. [In other words, the sum of the angles $\alpha+\beta+\gamma$ is **greater** than π.] If we substitute the formula $S = 4\pi r^2$ for the surface area of a sphere[34] of radius r then we obtain the exact formula

$$\triangle_{\alpha,\beta,\gamma} = \frac{(\alpha+\beta+\gamma-\pi)}{4\pi}\cdot 4\pi r^2 = (\alpha+\beta+\gamma-\pi)\cdot r^2$$

as desired. □

Let's kick the tires to see if this formula makes sense.

Check 1. Suppose you are standing at the North Pole with two friends. Your friends set off at right angles and each of them walks in a straight line until they hit the equator. Since they are both walking straight towards the equator, they will each hit it at a right angle. When this happens you and your friends will be at the three vertices of a geodesic triangle with **three right angles**:



[34] We'll prove this formula later.

Recall that a right angle equals $\pi/2$ in radian measure. If S is the surface area of the Earth then Harriot's formula tells us that the area of the big triangle is

$$\begin{aligned}\frac{\alpha+\beta+\gamma-\pi}{4\pi}\cdot S &= \frac{\pi/2+\pi/2+\pi/2-\pi}{4\pi}\cdot S\\ &= \frac{\pi/2}{4\pi}\cdot S\\ &= \frac{1}{8}\cdot S.\end{aligned}$$

And this **makes sense** because if your friends kept walking and returned to you at the North Pole, then their paths together with the equator would divide the surface of the Earth into 8 equal pieces. //

Check 2. How does the formula compare to our knowledge about the flat plane? Let's perform a thought experiment. Suppose that we have a triangle with interior angles α, β, γ on a sphere of radius r. Now imagine that we hold the area of the triangle **constant** while we slowly let the radius grow to infinity. (The angles will have to change as we do this.) In this case Harriot's formula says that

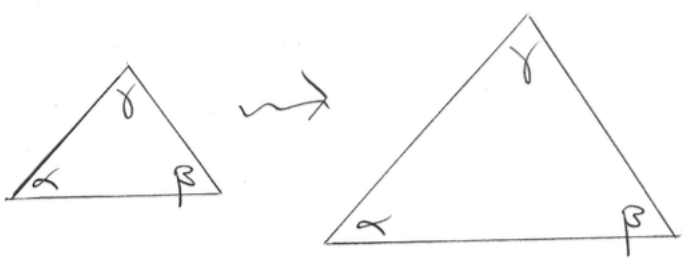
$$(\alpha+\beta+\gamma-\pi)\cdot r^2 = \text{constant}.$$

As the factor r^2 **grows**, the factor $(\alpha+\beta+\gamma-\pi)$ must **shrink** in order to keep the product constant. As the radius approaches infinity ($r\to\infty$) then we must have

$$(\alpha+\beta+\gamma-\pi)\to 0 \qquad \text{or} \qquad (\alpha+\beta+\gamma)\to\pi.$$

In other words, we conclude that a triangle on a sphere of "infinite radius" (i.e., on a flat plane) must have interior angles that sum to π (i.e., 180°). This agrees with our knowledge of Euclidean geometry. //

So the formula checks out. Now let's think about the consequences. Harriot's formula says that the area of a spherical triangle is determined by its angles. This is **very different** from the situation in Euclidean geometry, where we can "dilate" a triangle without changing its angles:



In Euclidean geometry we say that two triangles are *similar* when they have the same angles, and *congruent* when they have the same same side lengths. Euclid proved in Proposition I.8 that congruent triangles are necessarily similar [this is the side-side-side criterion for similarity] but similar triangles need not be congruent.

In contrast, the previous theorem can be extended to show that if two geodesic triangles on a sphere have the same angles then, in addition to having the same area, they **must also have the same side lengths**. In other words:

Similar triangles on a sphere are necessarily congruent.

This strange situation has everything to do with curvature. Indeed, recall that the surface of a sphere of radius r has constant Gaussian curvature $\kappa = 1/r^2 > 0$. This allows us to rewrite Thomas Harriot's formula as follows:

$$(\text{area of a triangle}) = (\text{angle excess}) \cdot r^2 = (\text{angle excess}) \cdot \frac{1}{\kappa} = \frac{(\text{angle excess})}{(\text{curvature})}.$$

And it turns out that this formula relating curvature to angles in a triangle can be generalized to arbitrary surfaces. This theorem was known to Gauss and it was first published by the French mathematician Pierre Ossian Bonnet in 1848.

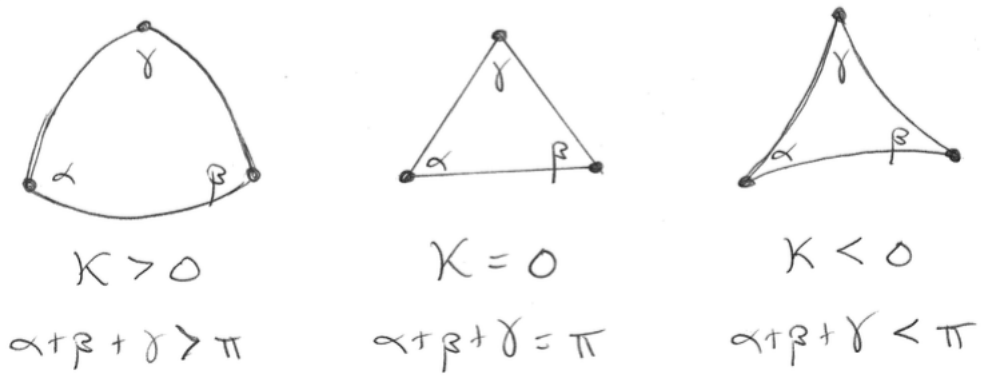
The Gauss-Bonnet Theorem (1848). Let $\triangle_{\alpha,\beta,\gamma}$ be a geodesic triangle on a smooth surface, with interior angles α, β, γ, and let κ be the **average value** of the Gaussian curvature inside the triangle. Then we have

$$(\text{area of } \triangle_{\alpha,\beta,\gamma}) = \frac{\alpha + \beta + \gamma - \pi}{\kappa}$$

$$(\text{area of a geodesic triangle}) = \frac{(\text{angle excess})}{(\text{average curvature inside the triangle})}$$

//

We can't apply this formula in the case of zero curvature because we can't divide by zero. However we can use a limiting argument as above to show that the angle excess approaches zero as κ goes to zero. The following picture summarizes our knowledge about geodesic triangles on a surface of **constant curvature** κ:

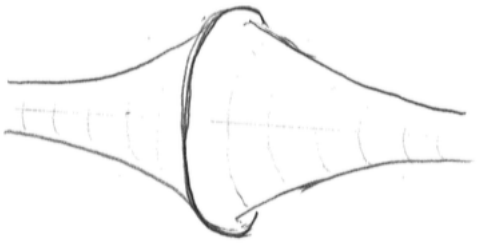


And what does this have to do with hyperbolic geometry? Recall that Beltrami and Klein proved that hyperbolic geometry is logically consistent[35] by finding a model of the hyperbolic plane that lives inside the Euclidean plane. The "lines" and "points" of this model are closely related to Euclidean "lines" and "points", however the "circles" in their model are squashed and in general the distance between points is distorted.

To end the chapter, I will outline a proof showing that it is **impossible** to find a distance-preserving Euclidean model of the hyperbolic plane:

- Giovanni Girolamo Saccheri (1733) showed that if Euclid's Parallel Postulate is false then the angles in any triangle sum to **less than** 180°. He regarded this result as absurd and so he considered it as a proof that the Parallel Postulate is **true**.
- Lobachevky (1830) and Bolyai (1832) both knew that if the Parallel Postulate is false then there exists some **negative constant** $\kappa < 0$ such that a triangle with interior angles α, β, γ has area
$$\frac{\pi - (\alpha + \beta + \gamma)}{\kappa}.$$
Since area is **positive** this means that the angle sum $\alpha + \beta + \gamma$ is **less than** π.
- Beltrami (1868) saw an analogy with Thomas Harriot's formula for the area of a spherical triangle. He suggested that the negative constant has the form $\kappa = -1/r^2$ and he said that the Bolyai-Lobchevsky geometry is a "pseudosphere" with imaginary radius $r \cdot \sqrt{-1}$.
- From the Gauss-Bonnet Theorem and the Theorema Egregium, we see that any Euclidean model of the hyperbolic plane must be a surface of **constant negative Gaussian curvature**. There does exist such a surface, called the *tractricoid*, but it has a sharp edge in the middle so it is not very satisfactory:

[35] At least, it is logically consistent as long is Euclidean geometry is logically consistent, which no one doubts.



- Finally, David Hilbert (1901) proved that every surface of constant negative Gaussian curvature must have a boundary or sharp edges. Thus there is no faithful Euclidean model of the hyperbolic plane.

Nevertheless, hyperbolic geometry **exists**.[36] Over time this example opened the floodgates to all kinds of "geometry" that can not be directly visualized. Some of it would shock you.

3 The Problem of Measurement

3.1 Pure and Applied Mathematics

The Greek word *geometria* literally means "land measurement". Indeed, the Greek historian Herodotus[37] (c. 484–425 BC) tells us that the Greeks learned geometry from the Egyptian art of measuring land. The following quote is taken from *The Histories* (c. 440 BC):

> This king also (they said) divided the country among all the Egyptians by giving each an equal parcel of land, and made this his source of revenue, assessing the payment of a yearly tax. And any man who was robbed by the river of part of his land could come to Sesostris and declare what had happened; then the king would send men to look into it and calculate the part by which the land was diminished, so that thereafter it should pay in proportion to the tax originally imposed. **From this, in my opinion, the Greeks learned the art of measuring land** [my emphasis]; the sunclock and the sundial, and the twelve divisions of the day, came to Hellas from Babylonia and not from Egypt.

Herodotus refers here to the ancient Babylonian and Egyptian civilizations, which left behind the earliest written evidence of matheamtics. These traditions were roughly contemporary (starting around 2000 BC) but we know more about Babylonian mathematics because their clay tablets were more durable than the Egyptian papyrus scrolls.

Most Babylonian and Egyptian sources are workbooks for solving practical computation problems. For example, one of the earliest Egyptian sources is the *Rhind Papyrus* (from before 1800 BC) which contains many geometric problems related to agriculture. The first geometric problem in the Rhind Papyrus reads as follows:

[36] To the detriment of Immanuel Kant's reputation.

[37] Not always a reliable source.

Problem 41. *Find the volume of a cylindrical grain silo with a diameter of* 9 *cubits and height of* 10 *cubits.*

The author then proceeds to give an algorithmic solution to the problem. In modern algebraic notation we can express their procedure with the formula

$$V = \left(d - \frac{d}{9}\right)^2 \cdot h,$$

where h is the height and d is the diameter of the cylinder. By applying this procedure to the values $d = 9$ cubits and $h = 10$ cubits they obtain the result

$$\begin{aligned} V &= 64 \text{ cubed cubits} \\ &= 900 \textit{ khar} \\ &= 4800 \textit{ hekat}. \end{aligned}$$

How accurate is this answer? If $r = d/2$ is the radius of the silo then the circular base of the silo has area πr^2 and the volume is given by the area of the base times the height:

$$V = \pi r^2 h.$$

By substituting $d = 2r$ into the Egyptian formula we obtain

$$V = \left(2r - \frac{2r}{9}\right)^2 \cdot h = \left(\frac{16r}{9}\right)^2 \cdot h = \left(\frac{16}{9}\right)^2 r^2 h.$$

We see that this expression has the correct form (constant)r^2h and that the Egyptians are using the value $(16/9)^2 = 266/81 = 3.160493827$ for the numerical constant $\pi \approx 3.14159$. It is not clear if the Egyptians knew or cared that this answer is not exact. From the form of the question we see that they are measuring volumes of wheat, presumably for the purposes of feeding people, and their solution is good enough that no one would have starved because of a mathematical error.

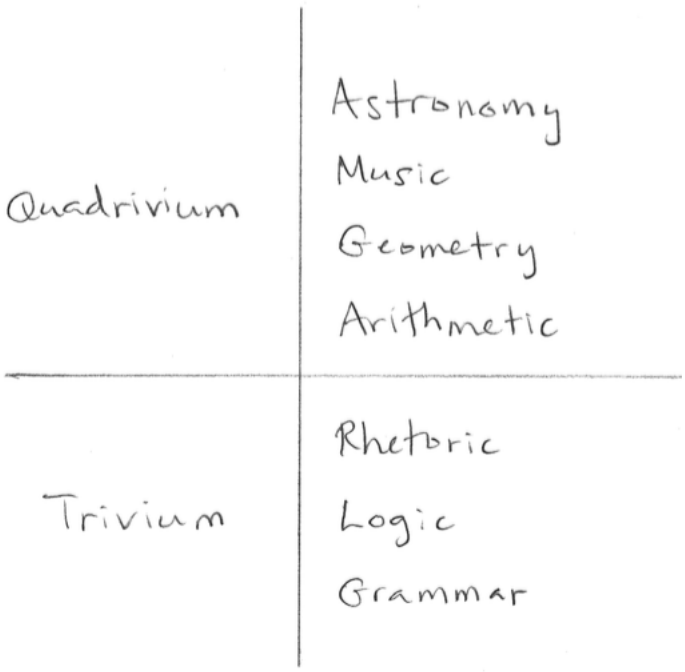
Now let's compare this to the Euclidean geometry of the Greeks. Euclid used the word *geometria* for the topic of the *The Elements*, but it's clear that Euclid's axiomatic geometry has nothing to do with land measurement. As I described in Chapter 1, the mathematicians of the Pythagorean tradition found a completely new purpose for mathematics that was spiritual and moral, rather than technological. I will refer to this new Pythagorean mathematics as "pure", as opposed to the earlier "applied" mathematics of the Babylonians and Egyptians.

Of course, applied mathematics continued to be practiced in Greece, but it was demoted to a secondary status below the pure mathematics of the philosophers. This distinction between pure and applied mathematics was stated most forcefully by Plato. In his dialogue *The Republic* (c. 380) the character Socrates and his companions discuss the characteristics of an ideal city (*eutopia*). A large part of the discussion centers on the education that the Guardians of the city should receive, which includes **10 years of mathematical training**.

The character Glaucon suggests that geometry will be useful for setting up military camps and troop formations, but Socrates responds that only a tiny amount of geometry is necessary for military purposes; instead, the Guardians should study advanced geometry because it is important for their moral development:

> What we have to consider is whether the greater and more advanced part of it tends to facilitate the apprehension of the idea of the Good. That tendency, we affirm, is to be found in all studies that force the soul to turn its vision round to the region where dwells the most blessed part of reality, which it is imperative that it should behold.

Thus, according to Plato, the true value of mathematics is that it teaches us what is good and true in the world (i.e., it "builds character") and thus it is ideal training for the leaders of society. This idea was taken up by the Romans with their concept of the "liberal arts" (*artes liberales*), which are the essential things that a free person needs to know. The Latin prose writer Martianus Capella (fl. 410–420 AD) divided the liberal arts into seven categories:



Despite its pagan origins, this system was taken up by the Catholic church and became the framework for the education system in Medieval Europe. The Trivium[38] of "verbal arts" was similar to our grade school and the more advanced Quadrivium of "numerical arts" was similar to our college curriculum.

Ever since the time of Plato, there has been some tension between the practice of mathematics

[38] origin of the word "trivial"

for technological reasons and the practice of mathematics for moral reasons. Indeed, after the word *geometria* was taken over by Euclidean abstract geometry, the word *metrika* (for "measurement") began to be used for the older, more computational, style of geometry.[39]

Because Euclidean geometry was disdainful of measurement and calculation it ignored some of the legitimate mathematical achievements of the Babylonians and Egyptians. In this chapter we will investigate the work that came after Euclid to reunite the pure and applied branches of geometry. The key figure in our story will be the Hellenistic (not Greek) scientist Archimedes of Syracuse (c. 287–212 BC).

In general I will refer to the project of reuniting pure and applied geometry as

the problem of measurement.

The goal here is to assign **numbers** to **geometric magnitudes** (such as lengths, angles, area and volume) so that one can solve geometric problems via arithmetic computations. This problem is quite difficult because it necessarily involves the concept of infinity; it was not really solved until Newton and Leibniz' invention of the *definite integral* in the 1660s. And even then the logical difficulties were formidable. The subject of *measure theory* was finally given an axiomatic foundation in the early 20th century by the French mathematicians René-Louis Baire, Émile Borel and Henri Lebesgue. It is still an active area of research.

3.2 Eudoxus' Theory of Proportion

To appreciate the difficulty of the problem of measurement we must recall from Chapter 1 the Pythagorean discovery that

the side and diagonal of a square are incommensurable.

The word *incommensurable* literally means "cannot be simultaneously measured". That is, if we assume that m and n are "numbers" (for the Greeks this means whole numbers) then there does not exist a common unit of measurement such that the diagonal of a square is m units long and the side is n units long. In modern notation we have no trouble saying that

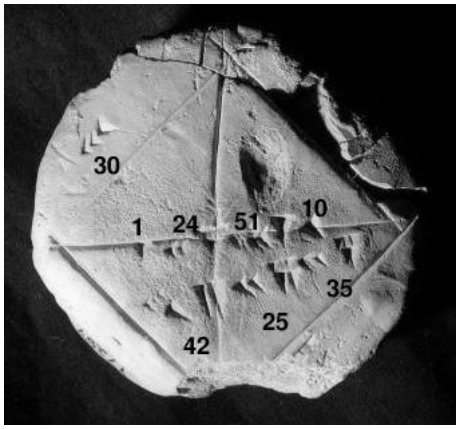
$$(\text{diagonal of a square}) = \sqrt{2} \cdot (\text{side length of a square})$$

but what does this really mean? In what sense are $\sqrt{2}$ and the side length "numbers" that can be multiplied together?

Here is a picture of a Babylonian clay tablet[40] dated from around 1700 BC. Its name is YBC7289 and it comes from the Yale Babylonian Collection:

[39] See for example the work *Metrika* (c. 60 AD) by Hero of Alexandria.

[40] borrowed from John Baez's *Azimuth* blog: `https://johncarlosbaez.wordpress.com/2011/12/02/babylon-and-the-square-root-of-2/`



We see here a picture of a square with a diagonal drawn. According to experts, the cuneiform writing on the diagonal stands for the number

$$1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3} = \frac{30547}{21600} \approx 1.41421,$$

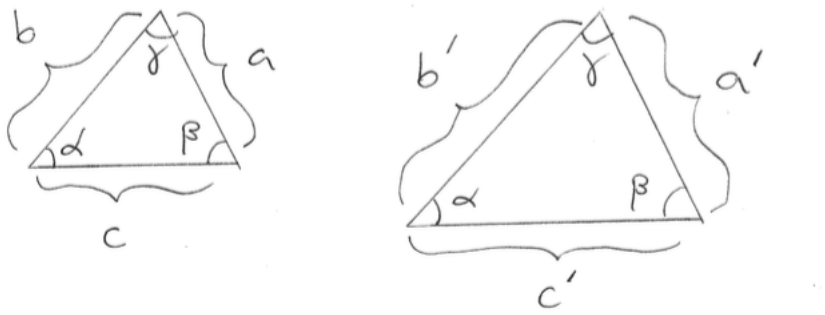
which is an impressively good approximation for the square root of 2. If this kind of technology was available in 1700 BC then the Pythagorean mathematicians over 1000 years later should have had no **practical** problem with the diagonal of a square. Their problem was **theoretical**. When they discovered that the square root of 2 is not a "number", this suggested that there is some mystery in the relationship between "numbers" and "geometric magnitudes".

Eudoxus of Cnidus (c. 309–337) was a student of Plato and he is regarded as the greatest of the classical Greek mathematicians.[41] Eudoxus knew that the ratio between some geometric magnitudes (such as the side and diagonal of a square) cannot be computed as a ratio of whole numbers. To get around this issue he developed an elaborate theory of *proportions*, which many people believe is faithfully preserved in Book V of Euclid's *Elements*. Before discussing Eudoxus' theory, let me show you the kind of problem that it is designed to solve.

Book VI of the *Elements* is all about "similar" geometric figures. The most famous proposition is the one about "similar triangles":

Proposition VI.4 (Similar Triangles are Proportional). Consider two triangles in the plane with the same interior angles α, β, γ and suppose that they have side lengths a, b, c and a', b', c' respectively:

[41] We will see that his geometric discoveries were later surpassed by Archimedes.



In this case I claim that the corresponding sides are "proportional", which in modern notation we can express as follows:

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

//

But let's think about this statement a bit. If the lengths a and b are commensurable then we can write $a/b = n/m$ where m and n are **whole numbers**. Similarly, if the lengths a' and b' are commensurable then we can write $a'/b' = n'/m'$ for some whole numbers $m' = n'$. Then the equation

$$\frac{n}{m} = \frac{n'}{m'}$$

can be expressed as the **equality of whole numbers** $mn' = m'n$. This is the idea of "cross-multiplication". But we know that many triangles have incommensurable lengths. For example, suppose that α is a right angle. Then results from Book I tell us that $b = c$ and the Pythagorean Theorem says that

$$\begin{aligned} a^2 &= b^2 + b^2 \\ a^2 &= 2b^2 \\ a^2/b^2 &= 2 \\ (a/b)^2 &= 2. \end{aligned}$$

In this case the ratios a/b and a'/b' can not be expressed in terms of whole numbers, so it is not clear what the equality $a/b = a'/b'$ should even mean. Nevertheless, we expect that Proposition VI.4 should still be true in the incommensurable case.

Eudoxus' theory of proportions is precisely what is needed to make this rigorous. Here are the relevant definitions recorded in Euclid's Book V:

> **Definition V.1.** A magnitude is a *part* of a magnitude, the less of the greater, when it measures the greater.

Definition V.2. The greater is a *multiple* of the less when it is measured by the less.

Definition V.3. A *ratio* is a sort of relation in respect of size between two magnitudes of the same kind.

Definition V.4. Magnitudes are said to *have a ratio* to one another which can, when multiplied, exceed one another.

Definition V.5. Magnitudes are said to be *in the same ratio*, the first to the second and the third to the fourth, when, if any equimultiples whatever are taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively taken in corresponding order.

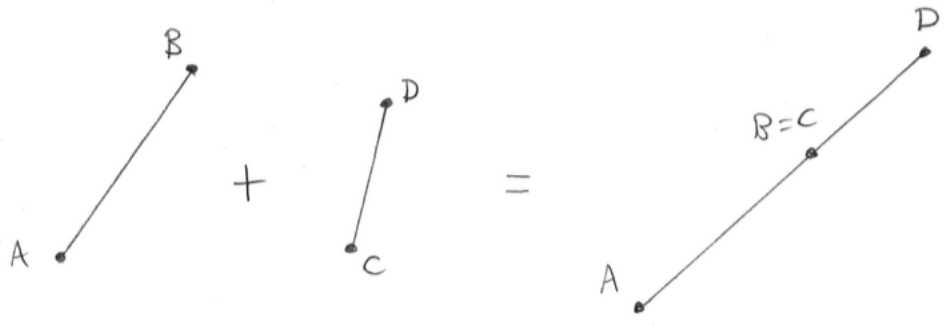
Definition V.6. Let magnitudes which have the same ratio be called *proportional*.

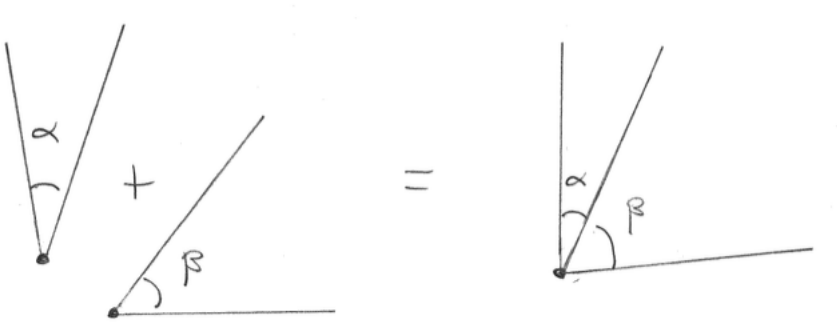
//

These definitions are quite obscure so let me try to express them in modern language.

Modern Explanation of Definitions V.1–V.6.

Definition V.1 refers to *magnitudes*. Euclid leaves this term undefined, but the idea is that a magnitude is anything that has "size" such as length, angle, area or volume. Magnitudes are very explicitly **not numbers** but they can potentially be measured by numbers. Magnitudes of the same kind can be "added" by sticking them together. For example, we can add line segments and angles as follows:





In general, however, **it makes no sense to add magnitudes of different kinds**. Given a magnitude X and a whole number n then we can add n identity copies of X together. In this case we use the notation

$$n \cdot X = \underbrace{X + X + \cdots + X}_{n \text{ times}}.$$

Thus, in some sense, is is possible to "multiply" a magnitude (which is not a number) by a whole number. However, **two magnitudes can not be multiplied together, even if they have the same kind**.[42]

Now consider two magnitudes X and Y of the same kind and a whole number n. If Y can be decomposed into n identical copies of X then we will write

$$n \cdot X = Y.$$

In this case Euclid says that X is a *part* of Y and Y is a *multiple* of Y. We could also think of X as some unit of measurement, in which case Euclid says that Y is *measured* by X. **This explains Definitions V.1 and V.2.**

Next Euclid wants to compare magnitudes. If X and Y are magnitudes of the same kind then I will write $X \subset Y$ to indicase that X is fully contained inside Y. Euclid says that the magnitudes X and Y *have a ratio* if there exist whole numbers m and n such that

$$m \cdot X \supset Y \qquad \text{and} \qquad X \subset n \cdot Y,$$

that is, if X fits inside some number of copies of Y and Y fits inside some number of copies of X.[43] If X and Y "have a ratio" then I will denote this ratio by

$$X : Y$$

and I will call it the ratio of X to Y. This ratio $X : Y$ is very explicitly **not a number**. However, in the special case of commensurable line segments with $m \cdot X = n \cdot Y$ then it is

[42] In Euclid the product of two line segments (which have length) is a rectangle (which has area), but the product of two general magnitudes (for example, the product of two angles) has no meaning.

[43] Euclid doesn't say whether every pair of magnitudes (of the same kind) "has a ratio". This assumption was explicitly made by Archimedes so it is often called the *axiom of Archimedes*.

correct to think of $X : Y$ as the same as the ratio $n : m$ of whole numbers. In modern language we might justify this idea by thinking of the ratio $X : Y$ as a "fraction X/Y" and then we might perform "cross-multplication":

$$m \cdot X = n \cdot Y \quad \Longleftrightarrow \quad \text{“}\frac{X}{Y} = \frac{n}{m}\text{”}.$$

However, we have to be careful with this because Euclid had no such concept of "fractions". For **incommensurable** line segments X and Y (such as the side and diagonal of a square) their ratio $X : Y$ is an abstract entity with no obvious relationship to whole numbers. **This completes the explanation of Definitions V.3 and V.4.**

Ratios of magnitudes (of the same kind) are **not numbers**; they can not be added or multiplied together. However, Euclid does need to say when two ratios are **equal**. This is the subject of **Definitions V.5 and V.6**. If $X_1 : Y_1$ and $X_2 : Y_2$ are commensurable ratios, i.e., if there exist whole numbers m_1, m_2, n_1, n_2 such that $m_1 \cdot X_1 = n_1 \cdot Y$ and $m_2 \cdot X_2 = n_2 \cdot Y_2$, then Euclid defines equality of ratios as follows:

$$X_1 : Y_1 = X_2 : Y_2 \quad \Longleftrightarrow \quad n_1 : m_1 = n_2 : m_2 \quad \Longleftrightarrow \quad m_1 n_2 = m_2 n_1.$$

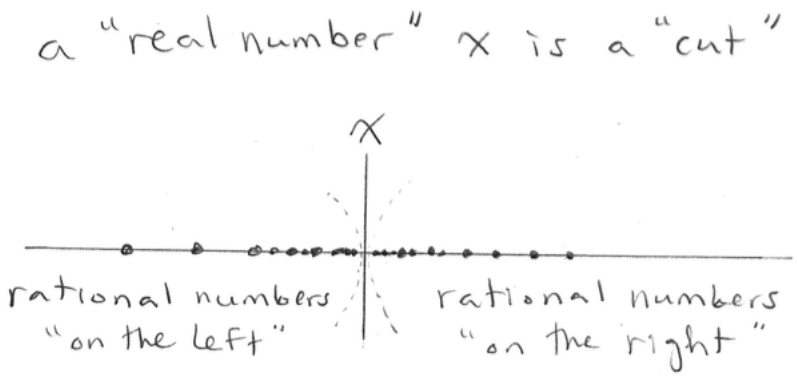
That is, the equality of commensurable ratios is defined by a certain equality of whole numbers. However, if $X_1 : Y_1$ and $X_2 : Y_2$ are **incommensurable** ratios then it is harder to say when they are equal. In this case **Definition V.5** says that ratios $X_1 : Y_1$ and $X_2 : Y_2$ are equal if **for all whole numbers m and n the following logical equivalences hold**:

$$\begin{aligned} m \cdot X_1 \subset n \cdot Y_1 \quad &\Longleftrightarrow \quad m \cdot X_2 \subset n \cdot Y_2 \\ m \cdot X_1 \supset n \cdot Y_1 \quad &\Longleftrightarrow \quad m \cdot X_2 \supset n \cdot Y_2 \end{aligned}$$

That's pretty confusing so let me translate it into today's language. I will temporatily think of a ratio $X : Y$ as a "fraction X/Y" and I will temporarily think of the relation $m \cdot X \subset n \cdot Y$ as an "inequality of fractions $X/Y < n/m$" by using "cross-multiplication". In this language, we can say that the two ratios $X_1 : Y_1$ and $X_2 : Y_2$ are equal precisely if **there do not exist whole numbers m and n such that**

$$\frac{X_1}{Y_1} < \frac{n}{m} < \frac{X_2}{Y_2} \qquad \text{or} \qquad \frac{X_1}{Y_1} > \frac{n}{m} > \frac{X_2}{Y_2}.$$

This definition is subtle because it involves the idea of "infinity": in order to check the equality of two ratios we need to show that **infinitely many whole numbers** satisfy a certain property. But this difficulty is unavoidable when dealing with incommensurables; in fact, Euclid's Definition V.5 has never really been improved on. The modern version was stated by Richard Dedekind (1813–1916) in his pamphlet on *Continuity and Irrational Numbers* (1872). Dedekind adapted Definition V.5 when he defined a "real number" (you can think of this as any point on the number line) as "cut" (German: *Schnitt*) that divides all of the rational numbers into "those on the left" and "those on the right" of the cut:



A consequence of this definition is that **two real numbers are equal precisely when there is no rational number between them**, because then they define the same cut. Thus we have the following dictionary between ancient Greek and modern mathematics:

Euclid (Eudoxus)	Dedekind
ratio	real number
commensurable ratio	rational number
incommensurable ratio	irrational number

The big difference between Euclid and Dedekind is that Dedekind viewed Euclid's "ratios" as "numbers" that can be added/subtracted and multiplied/divided, whereas Euclid viewed the "ratios" only as comparisons between magnitudes; **not** as magnitudes in themselves.

Finally, **Definition V.6** uses the word *proportional* (Greek: *anologon*) to refer to ratios that are equal in the sense of Definition V.5. This is why the subject of Book V is often referred to as Eudoxus' theory of *proportions*. //

These six definitions are the heart of Book V. The rest of the definitions and propositions are mostly boring; they just show that obvious properties of commensurable ratios, such as

$$X_1 : Y_1 = X_2 : Y_2 \quad \Longleftrightarrow \quad X_1 : X_2 = Y_1 : Y_2,$$

are still true for incommensurable ratios. Given the fact that the Babylonians could approximate $\sqrt{2}$ to five decimal places in 1700 BC, it might seem that Eudoxus' theory of proportion is unnecessarily abstract. However, we have to recall that the goal of Euclid's *Elements* is to rigorously **prove** the elementary facts of geometry, and from this point of view Eudoxus' theory is completely necessary.

Indeed, without the theory of proportion in Book V, Euclid couldn't even state the basic fact about proportionality of similar triangles. This is why he had to postpone the theory of similarity until Book VI. For example, here is the official statement of Proposition VI.4:

Consider two triangles with side lengths a, b, c and a', b', c'. If the corresponding angles are **equal** *then the corresponding side lengths are* **proportional** *as follows:*

$$a : a' = b : b' = c : c'.$$

Since most interesting triangles have incommensurable side lengths, we cannot even state this theorem without invoking the subtle Definition V.5. Most high schools teach the subject of similar triangles but they just skip over Book V and rely on the students' intuition about continuity and "real numbers".

The early users of the infinitesimal calculus (invented in the 1660s by Leibniz and Newton) also relied on their intuition. It wasn't until the 1800s that a rigorous theory of "real analysis" was developed that could finally bring the "applied mathematics" of Babylonian and Egyptian measurement together with the "pure mathematics" of Greek geometry. In this chapter we won't make it all the way to modern analysis but we will discuss some of the key developments along the way.

3.3 Archimedes and the Existence of π

We have seen that the Greeks before Euclid (primarily Eudoxus) developed an elaborate and abstract theory of *proportion* in order to deal with incommensurable ratios between geometric magnitudes. The theory worked well for measuring the lengths of straight line segments, and it was also successful for measuring the areas of some curved regions. However, Eudoxus' method completely failed to solve the following two problems:

1. Measure the length of the arc of a circle.
2. Measure the length of a general curved path.

The first of these is the problem of *trigonometry*. The Greeks before Euclid made very little progress on this problem. In fact, Aristotle (c. 384–322 BC) in his *Physics* explicitly stated that it is **impossible** to compare straight line motion with motion in a circle:

> The fact remains that if the motions are comparable, there will be a straight line equal to a circle. But the lines are not comparable; so neither are the motions.

Some progress on the problem was made by the post-Euclidean Hellnistic mathematicians Archimedes and Ptolemy. Further development occurred in India and the Islamic world and then the subject returned to Europe during the Renaissance, but the modern form of the subject didn't emerge until the development of Calculus in the 17th century. The word *trigonometria* (literally "triangle measurement") was coined by Bartholomeo Pitiscus in 1595. The modern notation for the subject (in terms of *sine*, *cosine*, *tangent*, etc.) was finally standardized by Leonhard Euler in his textbook *Introductio in analysin infinitorum* (1748) (Introduction to Analysis of the Infinite).

From the above discussion you might get the sense that trigonometry is part of Calculus, but this is not so. The subject can be developed in an elementary geometric way; it just happens that ideas from Calculus make the subject easier and more coherent. The second problem stated above, however, is completely impossible without the language of Calculus.

In this section I will show how the subject of trigonometry emerged from the following hard problem:

measure the circumference of a circle in terms of its diameter.

Let C and d denote the circumference and diameter of a given circle. In modern terms we denote the ratio C/d by the symbol π. We know that this is an "irrational number" which can be approximated by a decimal expansion $\pi = 3.14159\cdots$.[44] But how do we know that the "number" π even exists? When we use the symbol π we are implicitly assuming the following theorem.

Theorem (Existence of π). Consider two circles with circumferences C and C' and with diameters d and d'. Then the ratio of circumference to diameter is the same for both circles:

$$\frac{C}{d} = \frac{C'}{d'}.$$

In modern terms we think of this common ratio as a number and we call it π. //

It is surprisingly difficult to track down the history of this theorem.[45] The empirical truth of the statement must have been known to all ancient cilivizations. Indeed, the ancient Egyptian and Babylonian mathematicians were using reasonable approximations to π before 1800 BC. Therefore it is certain that the ancient Greeks knew of the fact $C/d = C'/d'$, and yet no mention of this result appears in Euclid's *Elements*. The absence of π in Euclid's *Elements* is puzzling to modern readers; I surmize that it was left out because Euclid was unable to **prove** the statement $C/d = C'/d'$. After all, it would have been out of character for Euclid to provide partial information on a ratio whose existence he could not prove.

My belief that Euclid knew but could not prove that $C/d = C'/d'$ is supported by the following theorem that he **did** prove.

Proposition XII.2. *Circles are to one another as the squares on their diameters.* //

In modern terms: Consider two circles with areas A and A' and with diameters d and d'. Then the ratio of area to diameter squared is the same for both circles:

$$\frac{A}{d^2} = \frac{A'}{(d')^2}.$$

[44] The use of the symbol π for this purpose was popularized by Leonhard Euler.

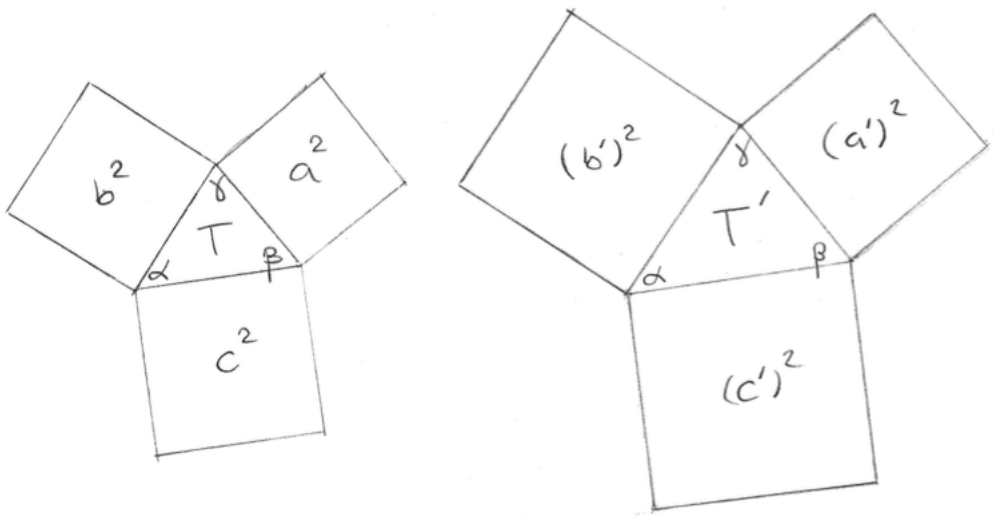
[45] This section borrows ideas from David Richeson's article *Circular Reasoning: Who First Proved That C/d is a Constant?*

Proof. The magnitudes A and d^2 both represent areas so they can be compared via Eudoxus' theory in Book V. The goal is to show that

$$A : d^2 = A' : (d')^2,$$

where the equality of ratios is understood in the sense of Definition V.5. I will present a sketch of the proof in modern language.

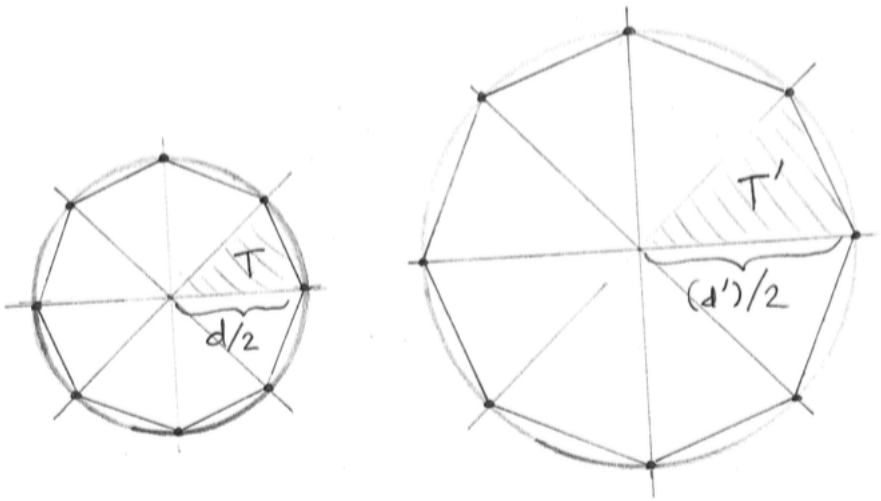
First of all, Euclid had shown the following result in Proposition VI.19. Consider two similar triangles with side lengths a, b, c and a', b', c', respectively, and construct the squares on their sides:



If the areas of the triangles are T and T', then we have

$$T : T' = a^2 : (a')^2 = b^2 : (b')^2 = c^2 : (c')^2.$$

Now consider a regular 2^n-gon inscribed in each circle. Here is the picture for $n = 3$:



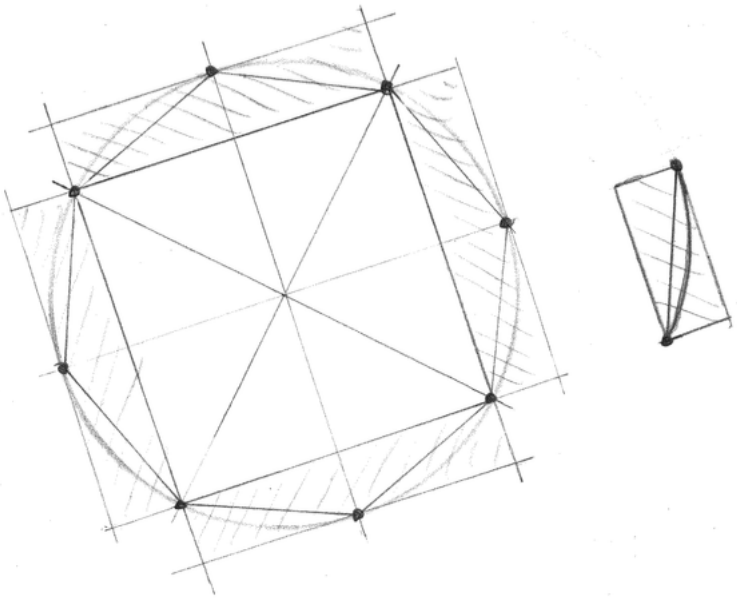
Let P_n and P'_n be the areas of the two 2^n-gons and decompose each into 2^n triangles as in the picture. Let T and T' denote the areas of these triangles, as in the figure. Since the triangles have the same angles we can use Proposition VI.19 from above and an obvious property of ratios [Euclid's Proposition V.15 says that $X : Y = m \cdot X : m \cdot Y$ for any m] to show that

$$\begin{aligned} P_n : P'_n &= 2^n \cdot T : 2^n \cdot T' \\ &= T : T' && \text{Prop V.15} \\ &= (d/2)^2 : (d'/2)^2 && \text{Prop VI.19} \\ &= d^2 : (d')^2. && \text{Prop V.15} \end{aligned}$$

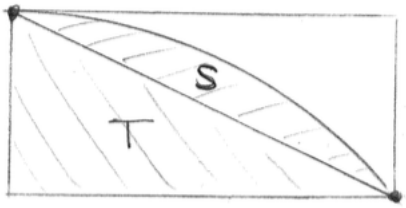
For the rest of the proof, the idea is to let "n approach ∞" and observe that "the ratio of polygon areas $P_n : P'_n$ approaches the ratio of circle areas $A : A'$". Since we know that $P_n : P'_n = d^2 : (d')^2$ for any n, it should follow from this that $A : A' = d^2 : (d')^2$ and hence we obtain $A : d^2 = A' : (d')^2$ as desired. [The obvious fact that $X_1 : Y_1 = X_2 : Y_2$ implies $X_1 : X_2 = Y_1 : Y_2$ is Euclid's Corollary V.7.]

If you have taken Calculus then you have probably seen this kind of "continuity argument" used in a non-rigorous way. To make the argument rigorous Euclid needed to refer back to Eudoxus' Definition V.5 on the proportionality of ratios. Here's how he did it.

First he observed that the difference $(A - P_n)$ of the circle area A minus the 2^n-gon area P_n can be made as small as we want by taking n large enough. For example, consider the following picture of the polygons P_2 and P_3 inscribed in the circle with area A:



The differences $(A - P_2)$ and $(A - P_3)$ are contained inside the four shaded rectangles shown and in fact it suffices to look at just half of one of these rectangles. Here is a zoomed-in version of the relevant rectangle with the areas S and T of a sector and a triangle shaded:



In the general case there will be 2^{n+1} such rectangles, thus we can express the area differences $(A - P_{n+1})$ and $(A - P_n)$ by the following formulas:

$$(A - P_{n+1}) = 2^{n+1} \cdot S \qquad \text{and} \qquad (A - P_n) = 2^{n+1} \cdot (S + T).$$

Now observe from the picture that we have $S < T$ because the sector S is completely contained in the upper triangle which is congruent to T (and hence has the same area as T). It follows from this that

$$\begin{aligned} (A - P_n) &= 2^{n+1} \cdot (S + T) \\ &> 2^{n+1} \cdot (S + S) \qquad\qquad \text{because } T > S \end{aligned}$$

$$= 2 \cdot 2^{n+1} \cdot S$$
$$= 2 \cdot (A - P_{n+1})$$

and hence

$$2 \cdot (A - P_{n+1}) < (A - P_n)$$
$$(A - P_{n+1}) < (A - P_n)/2.$$

In other words, if we increase n by 1 then the new difference $(A - P_{n+1})$ is less than half of the previous difference $(A - P_n)$. It now follows from Euclid's Proposition X.1 that *we can make the difference* $(A - P_n)$ *smaller than any fixed area by taking* n *sufficiently large.*

To complete the proof, Euclid assumes that there exists some area X with the property

$$X : A' = d^2 : (d')^2$$

and then he uses the method of **double contradiction** to show that this area must satisfy $X = A$. For the first contradiction, assume that $X < A$. Then in modern notation we can write $(A - X) > 0$ and from the previous result we can choose some n large enough so that

$$(A - X) > (A - P_n) > 0.$$

From this it follows that $P_n > X$. But we also know that $A' > P'_n$ because the polygon is contained completely inside the circle. Then from an obvious property of ratios [Proposition V.8 says that if $X < Y$ then we have $X : Z < Y : Z$ and $Z : X > Z : Y$ for all Z] and from our previous result $P_n : P'_n = d^2 : (d')^2$ we see that

$$\begin{aligned} d^2 : (d')^2 &= P_n : P'_n \\ &> X : P'_n && \text{because } P_n > X \\ &> X : A' && \text{because } A' > P'_n \\ &= d^2 : (d')^2. \end{aligned}$$

This is a contradiction because it says that the ratio $d^2 : (d')^2$ is strictly greater than itself. Since the assumption $X < A$ leads to a contradiction, it must be that X is **not smaller** than A. Euclid then gives a similar proof by contradiction to show that X is **not greater** than A. [That's why it's called the method of "double contradiction".] Finally, since X is not less than A and not greater than A, it follows that $X = A$ and we conclude that $A : A' = X : A' = d^2 : (d')^2$ as desired. □

Let's think about this result in modern terms. If we write $d = 2r$ and $d' = 2r'$ where r and r' are the radii of the two circles, then the theorem says that

$$A/d^2 = A'/(d')^2$$
$$A/(4r^2) = A'/(4(r')^2)$$
$$A/r^2 = A'/(r')^2.$$

Then since the ratio A/r^2 is the same for any two circles we can give it a special name. Let's call it

$$\pi' = \frac{A}{r^2}.$$

In other words, we have the formula $A = \pi' r^2$ for the area of a circle. These days we all know that $A = \pi r^2$ so it must be the case that the constant ratio $\pi' = A/r^2$ defined in terms of areas is **equal** to the constant ratio $\pi = C/d$ defined in terms of lengths, but Euclid makes no mention of this. It is possible that Euclid knew the formula $\pi = \pi'$ as an empirical fact but that he left it out of the *Elements* because he couldn't prove it rigorously. It is also possible[46] that the fact $\pi' = \pi$ was completely unknown in Euclid's time.

The absence of π in the *Elements* leaves us with two questions:

(1) Who first stated and proved that $C/d = \pi = \pi' = A/r^2$ for all circles?

(2) Who first stated and proved that $\pi = C/d$ is a constant for all circles?

The answer to the first question is: Archimedes. The answer to the second question is: probably Archimedes, but it's not explicitly written in any of his surviving works. The earliest known explicit statement and proof of (2) appears in a 9th century work called *Kitab fi ma'rifat misahat al-ashkal al-basita wa al-kuriya* (c. 850 AD) (The Measurement of Plane and Spherical Figures). It was written by the three Banu Musa brothers who were mathematicians and astronomers working at the House of Widsom in Baghdad.[47]

Their book was largely a commentary on Archimedes' work, which had recently been translated into Arabic by Thabit ibn Qurra (836–901 AD). In particular, they derived (2) as a direct consequence of Archimedes' theorem (1) and Euclid's Proposition XII.2.

Proof that π Exists. Archimedes proved[48] that for any circle we have

$$\frac{C}{d} = \frac{A}{r^2},$$

where C, d, A, r are the circumference, diameter, area and radius, respectively. Now recall from Euclid's Proposition XII.2 above that for any two circles we have

$$\frac{A}{r^2} = \frac{A'}{(r')^2}.$$

By putting these two equations together, it follows that for any two circles we have

$$\frac{C}{d} = \frac{A}{r^2} = \frac{A'}{(r')^2} = \frac{C'}{d'},$$

and hence the ratio C/d is a universal constant. □

[46] but I find it unlikely

[47] The House of Wisdom was the center of mathematical knowledge during the Islamic Golden Age, just as the Library of Alexandria was the center of mathematics during the Hellenistic period.

[48] see below for the proof

To complete the proof that π exists, it remains to show that we have $C/d = A/r^2$. This was the great contribution of Archimedes, to which we now turn.

Archimedes of Syracuse (c. 287–212 BC) is regarded as the greatest mathematician of antiquity, and possibly the greatest mathematician of all time. He was born in Syracuse, Sicily, approximately two generations after Euclid. He studied at the Library of Alexandria so he would have been intimately familiary with Euclid's work. But then he chose to move back to Sicily where he spent the rest of his life. Unlike the pure geometers of the Euclidean tradition, Archimedes was deeply interested in applications of his work to practical problems. He pioneered the use of screws, pulleys and levers, and he was renowned for the military technology that he developed for King Hiero II of Syracuse. His greatest discovery in physics is called *Archimedes' princple*, which he described in his work *On Floating Bodies* as follows:

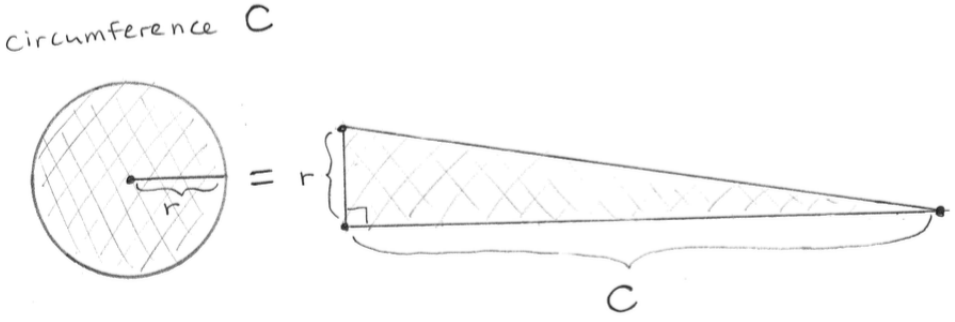
> Any object, wholly or partially immersed in a fluid, is buoyed up by a force equal to the weight of the fluid displaced by the object.

Legend says that he discovered this principle while in the bath and that he was so pleased with the discovery that he ran naked through the streets of Syracuse shouting "Eureka!" (I have found it!).

Archimedes' most famous mathematical works are the *Measurement of a Circle* and *On the Sphere and Cylinder*, both written around 225 BC. Scholars disagree on which came first, so they are usually regarded as a single work. The great achievement of this work is that it completely solves the measurement problem for spheres and circles, something pre-Euclidean mathematicians had failed to do. Archimdes' theorem on the area of a circle is so famous that we can hardly imagine a time before it was known.

Theorem (*Measurement of a Circle*, Proposition 1). *The area of any circle is equal to a right-angled triangle in which one of the sides about the right angle is equal to the radius, and the other to the circumference, of the circle.*

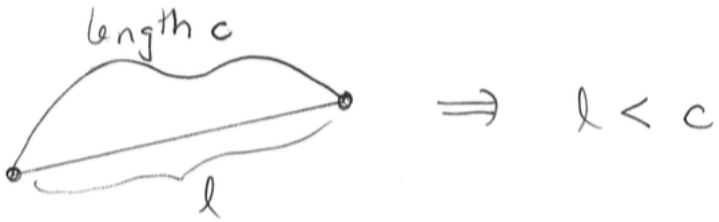
In modern terms, consider a circle with circumference C and radius r. Then the area A of the circle is equal to the area of a right triangle with side lengths C and r as in the following picture:



//

Archimedes realized that it was likely impossible to prove this theorem on the basis of Euclid's *Elements*, but he didn't let that hold him back. Perhaps because of his geographic distance from Alexandria he was bold enough to suggest new axioms on top of Euclid's 5 Postulates and 5 Common Notions. He stated these axioms explicitly in *On the Sphere and Cylinder* and he used them implicitly in *Measurement of a Circle*. Here are the two axioms that are needed for the proof of Archimedes' Proposition 1.

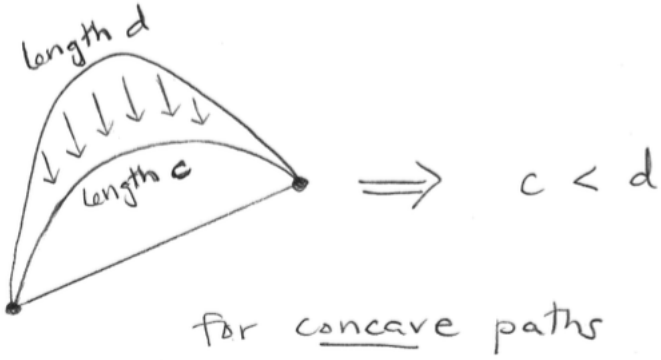
Archimedes Postulate 1: *That among lines which have the same limits the straight line is the smallest.*



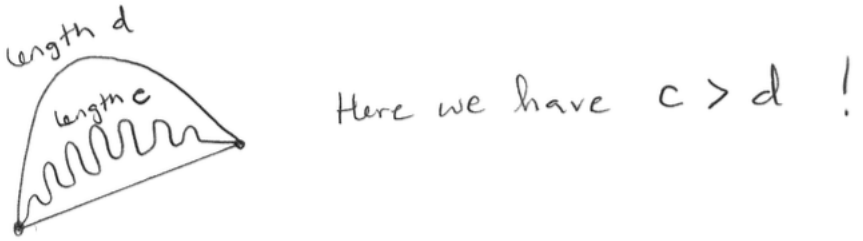
In modern terms, this axioms says that a straight line is the shortest path between two points.[49] The next axiom says that paths become shorter as they approach a straight line.

Archimedes Postulate 2: *And, among the other lines (if, being in a plane, they have the same limits), that such lines are unequal, when they are both concave in the same direction and either one of them is whole contained by the other and by the straight line having the same limits as itself, or some is contained, and some it has as common, and the contained is smaller.*

[49] In other words, straight lines are *geodesic paths* in Euclidean geometry. We already assumed this fact in Section 2.4.



In modern terms, we have two concave paths on the same side of a straight line segment. (The term *concave* means that the paths don't wobble back and forth.) If one curve is completely between the line and the other curve, then this path is shorter. The following picture illustrates why the "concave" condition is necessary:



//

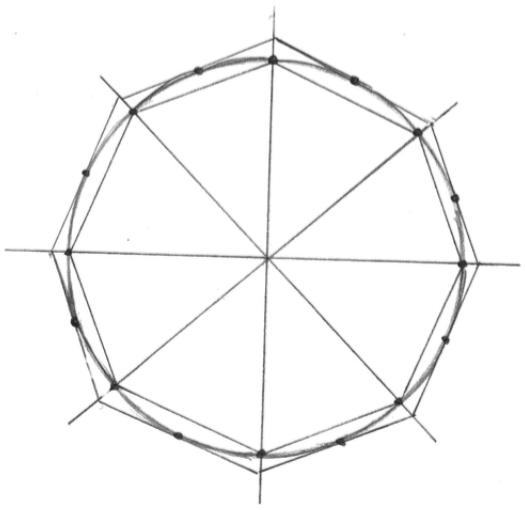
These are reasonable axioms because they are self-evidently true.[50] Both of these facts can be **proved** from Euclid's *Elements* when the paths are *piecewise-linear*, i.e., composed of line segments glued together. Archimedes' axioms are only necessary when dealing with curved paths. We will apply them to arcs of a circle.

Proof of Archimedes' Proposition 1. Consider a circle and denote its area, circumference and radius by A, C and r, respectively. The goal is to show that

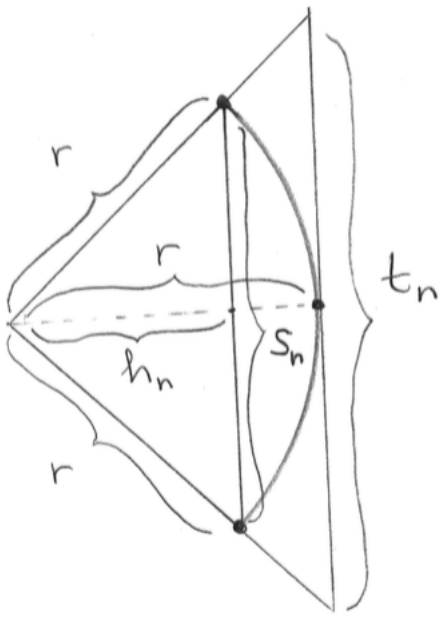
$$A = \frac{1}{2}Cr.$$

For this purpose Archimedes considered two regular 2^n-gons, one inscribed in the circle and one circumscribed around it. Here is a picture for $n = 3$:

[50] At least for short paths in a locally-flat surface. For long paths in a curved surface, who knows?



We will denote the inner 2^n-gon by P_n and the outer 2^n-gon by Q_n. Now we want to relate P_n and Q_n to the area and circumference of the circle. Here is a zoomed-in picture of $1/2^n$-th of the circle (the angle is exaggerated to make it more readable):



In this picture I have labeled the side lengths of the polygons P_n and Q_n by s_n and t_n,

respectively, so that the perimeters are given by

$$\text{perim}(P_n) = 2^n \cdot s_n \qquad \text{and} \qquad \text{perim}(Q_n) = 2^n \cdot t_n.$$

Furthermore, note that each of P_n and Q_n is composed of 2^n identical triangles. Thus by using the formula

$$(\text{area of a triangle}) = \frac{1}{2} \cdot (\text{base})(\text{height})$$

we see that the areas of the polygons P_n and Q_n are given by

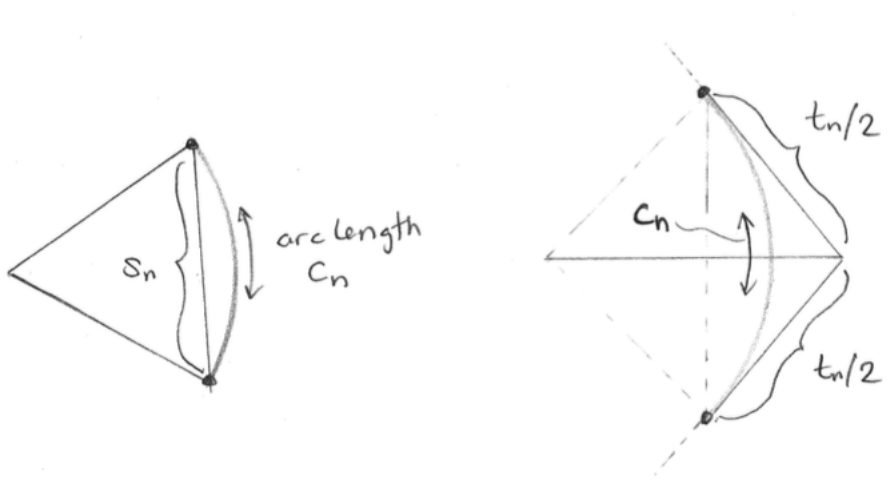
$$\begin{array}{ll} \text{area}(P_n) = 2^n \cdot \frac{1}{2} h_n s_n & \text{area}(Q_n) = 2^n \cdot \frac{1}{2} r t_n \\ \text{area}(P_n) = \frac{1}{2} h_n (2^n \cdot s_n) \qquad \text{and} & \text{area}(Q_n) = \frac{1}{2} r (2^n \cdot t_n) \\ \text{area}(P_n) = \frac{1}{2} h_n \cdot \text{perim}(P_n) & \text{area}(Q_n) = \frac{1}{2} r \cdot \text{perim}(Q_n) \end{array}$$

where h_n is the "height" shown in the diagram.

The rest of the proof will proceed by a **double contradiction**, exactly as in Euclid's Proposition XII.2. That is, we will show that each of the assumptions $A < \frac{1}{2}Cr$ and $A > \frac{1}{2}Cr$ leads to a contradiction, and so it must be the case that $A = \frac{1}{2}Cr$ as desired. To obtain the contradictions, however, we will first need to use Archimedes' Postulates 1 and 2 to prove the following reasonable fact:

$$\text{perim}(P_n) < C < \text{perim}(Q_n).$$

To prove these inequalities, let c_n denote the length of $1/2^n$-th of the circumference and consider the following pictures:



Then applying Postulate 1 to the picture on the left gives

$$\begin{array}{ll} s_n < c_n & \text{Postulate 1} \\ 2^n \cdot s_n < 2^n \cdot c_n & \end{array}$$

$$\text{perim}(P_n) < C$$

and applying Postulate 2 to the picture on the right gives

$$\begin{aligned} c_n &< t_n/2 + t_n/2 && \text{Postulate 2} \\ c_n &< t_n \\ 2^n \cdot c_n &< 2^n \cdot t_n \\ C &< \text{perim}(Q_n) \end{aligned}$$

as desired.

Finally, let us assume for contradiction that $A > \frac{1}{2}Cr$, so that

$$\left(A - \frac{1}{2}Cr\right) > 0.$$

It was shown in the proof of Euclid's Proposition XII.2 above that by choosing n large enough we can make the area $A - \text{area}(P_n)$ smaller than any given area. In particular, we can choose some n such that

$$\left(A - \frac{1}{2}Cr\right) > (A - \text{area}(P_n)) > 0,$$

and it follows from this that

$$\frac{1}{2}Cr < \text{area}(P_n). \qquad (*)$$

On the other hand, we showed above that $\text{area}(P_n) = \frac{1}{2}h_n \cdot \text{perim}(P_n)$ and $\text{perim}(P_n) < C$. Since h_n and r are two sides of a triangle it also follows from Euclid's Proposition I.20 that $h_n < r$. Putting these facts together gives

$$\text{area}(P_n) = \frac{1}{2}h_n \cdot \text{perim}(P_n) < \frac{1}{2}h_n C < \frac{1}{2}rC,$$

which **contradicts** the inequality $(*)$. We conclude from this contradiction that the inequality $A > \frac{1}{2}Cr$ is impossible. A similar argument (omitted) shows that the inequality $A < \frac{1}{2}Cr$ is also impossible. Therefore it must be the case that $A = \frac{1}{2}Cr$. □

This completes Archimedes' proof of the immortal formula $A = \pi r^2$, where π is defined as the ratio C/d. Putting together Archimedes' Proposition 1 and Euclid's Proposition XII.2 then yields a proof that $C/d = C'/d'$ for all circles, and hence the universal constant π **exists**. For whatever reason, Archimedes did not include a statement of this fact in the *Measurement of a Circle*, however he did use the same method of proof (by inscribing and circumscribing regular polygons) to compute the following extremely accurate bounds on the value of π.

Theorem (*Measurement of the Circle*, Proposition 3). *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 modern language:

$$\begin{array}{ccccc} 223/71 & < & C/d & < & 22/7 \\ 3.1407 & < & C/d & < & 3.1429 \end{array}$$

This proposition indicates that Archimedes was aware of the existence of π. It is possible that he stated this explicitly in some work that was lost, or perhaps he felt that the result was not important enough to write down. It is also possible that the prejudice against comparing straight lines and curved paths (as shown in the above quote from Aristotle) was too strong to allow Archimedes to claim a rigorous theorem about the ratio C/d.

Indeed, this prejudice persisted for thousands of years. In the revolutionary work *Géométrie* (1637) in which René Descartes introduced the idea of *(Des)cartesian coordinates* into geometry, he also made the following surprising statement:

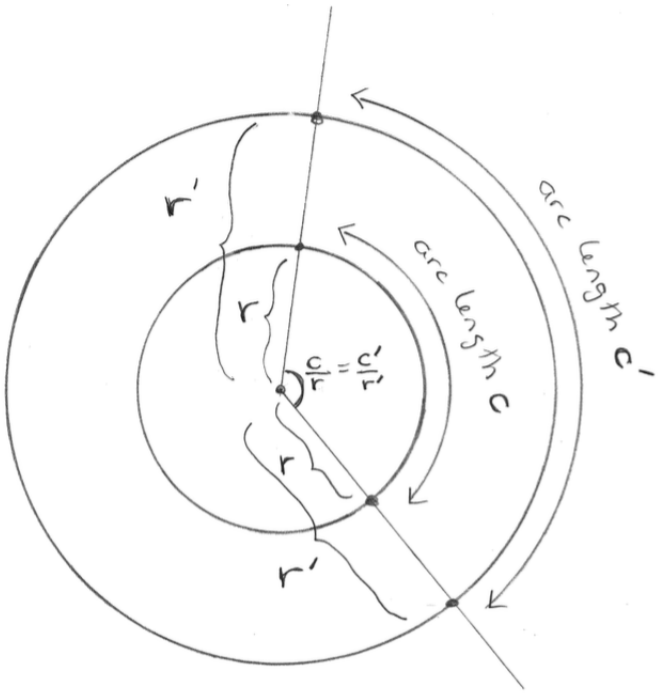
> Geometry should not include lines that are like strings, in that they are sometimes straight and sometimes curved, since the ratios between straight and curved lines are not known, and I believe cannot be discovered by human minds, and therefore no conclusion based upon such ratios can be accepted as rigorous and exact.[51]

3.4 Trigonometry is Hard

With hindsight we can view the theorem on the existence of π as an important step towards the modern subject of *trigonometry*. I mentioned above that the word *trigonometria* (coined by Bartholomeo Pitiscus in 1595) literally means "triangle measurement", but it is mathematically and historically more correct to think of the subject of trigonometry in terms of "circle measurement". In particular, the existence of π gives us a clever way to measure angles.

Idea of Radian Measure. Consider a fixed angle and draw any two concentric cirles with radii r and r' centered on the angle. Now let c and c' denote the lengths of the circular arcs cut out by the angle, as in the following picture:

[51] This quote and the Archimedes quote are taken from Richeson's *Circular Reasonong: Who First Proved that C/d is a Constant?*



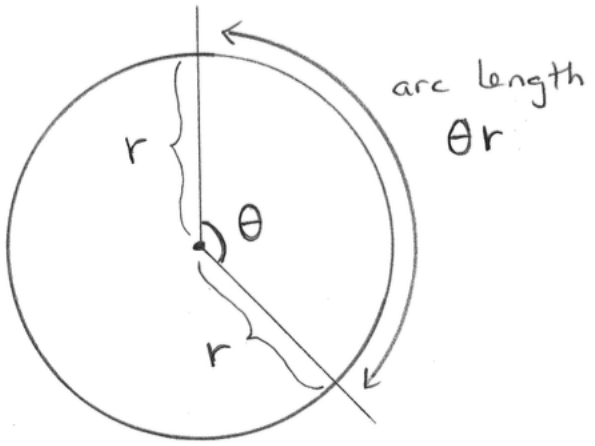
As a consequence of the theorem on the existence of π we have

$$\frac{c}{r} = \frac{c'}{r'}.$$

The idea of *radian measure* is to use this common ratio $c/r = c'/r'$ as a **measure of the angle**. In the extreme case that the angle fills up the whole circle then c is equal to the whole circumference $C = 2\pi r$ and hence the angle has measure

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

From this we see that measure of the angle can take any value between 0 and 2π. Conversely, for any number $0 \leqslant \theta \leqslant 2\pi$ we see that the angle of radian measure θ cuts out an arc length of θr from the circle of radius r:



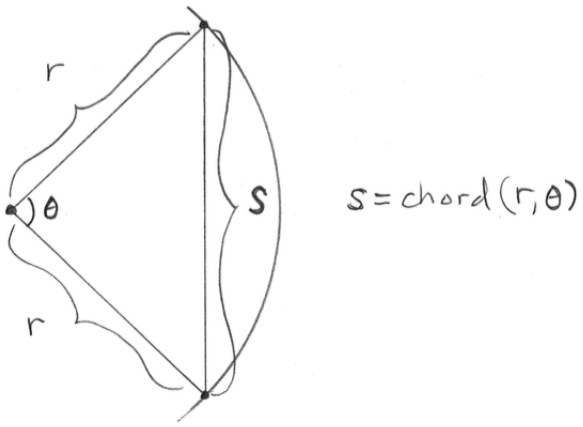
In other words, the circular arc cut out by the angle of radian measure θ has arc length equal to θ radii. This explains the terminology. //

This notation illustrates that

the problem of measuring an angle is equivalent to the problem of measuring the length of a circular arc.

And it is based on the fact that we know the relationship between the full circumference and the radius. But the radius is not the **only** straight line segment associated to a given angle/arc. The fundamental problem of trigonometry is to find an explicit relationship between a given angle/arc and the *chord of the circle* that it subtends.

The Fundamental Problem of Trigonometry. Consider a chord in a circle of radius r. Suppose that the chord has length s and subtends an angle θ as in the following picture:



Consider the isoceles triangle with side lengths r, r, s. Because of the *side-angle-side criterion* for congruence of triangles (which is Euclid's Proposition I.4) we know that the chord length s is uniquely determined by the radius r and the angle θ. In other words, we can think of s as a **function** of r and θ:

$$s = \text{chord}(r, \theta).$$

The fundamental problem of trigonometry is to **compute** this function. That is, for a fixed radius r, the problem is to compute s from θ or to compute θ from s. //

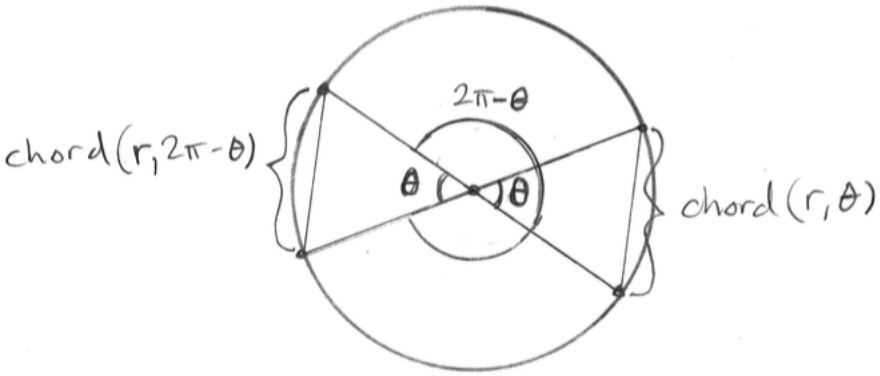
There are certain specific values of θ and s for which we know that answer. For example, if $\theta = 0$ or $\theta = 2\pi$ then the length of the chord is zero:

$$\boxed{\text{chord}(r, 0) = 0 \qquad \text{and} \qquad \text{chord}(r, 2\pi) = 0.}$$

In fact, we can restrict our attention to angles less than a straight line ($0 \leqslant \theta \leqslant \pi$) since for any angle θ we have

$$\boxed{\text{chord}(r, \theta) = \text{chord}(r, 2\pi - \theta).}$$

Proof: The triangles in the following picture are congruent by the side-angle-side criterion:

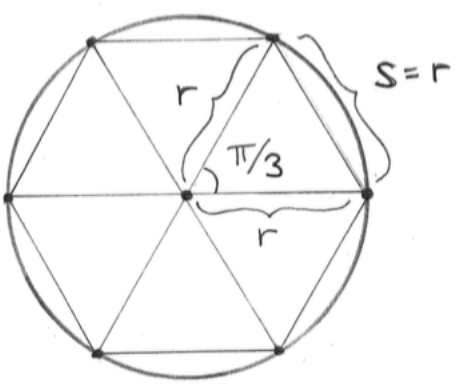


□

We also know that the angle $\theta = (2\pi)/6 = \pi/3$ corresponds to the chord length $s = r$:

$$\boxed{\text{chord}(r, \pi/3) = r.}$$

Proof: Consider a regular hexagon inscribed in the circle of radius r:



Since the hexagon is built out of six equilateral triangles we see that each of the sides of the hexagon has length r. □

[Remark: Using Archimedes' Postulate 1 (that the straight line is the shortest distance between two points) this argument also shows that

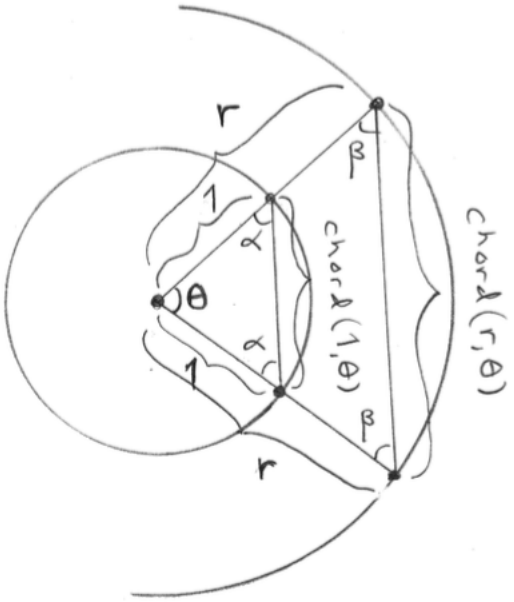
$$\begin{aligned}(\text{length of chord}) &< (\text{length of arc})\\ s &< (\pi/3)r\\ r &< (\pi/3)r\\ 1 &< \pi/3\\ 3 &< \pi,\end{aligned}$$

which gives us a crude lower bound for π.]

Finally, let me note that the radius is mostly irrelevant to the chord function. Indeed, if we can compute chord lengths for the "unit circle" of radius 1 then we obtain the chord lengths for any circle of radius r:

$$\boxed{\text{chord}(r,\theta) = r \cdot \text{chord}(1,\theta).}$$

Proof: Consider a fixed angle of θ in concentric circles of radius 1 and r. This defines two isoceles triangles with interior angles θ, α, α and θ, β, β as in the following picture:



The fact that the triangles are isoceles comes from the *Pons Asinorum* (Euclid's Proposition I.5). Then from the fact that the interior angles sum to 180° (or π radius) we conclude that

$$\begin{aligned} \theta + \alpha + \alpha &= \theta + \beta + \beta \\ 2\alpha &= 2\beta \\ \alpha &= \beta. \end{aligned}$$

In other words, the two triangles are similar (they have the same angles). Finally, it follows from Euclid's Proposition VI.4 (proportionality of similar triangles) that

$$\begin{aligned} \frac{1}{r} &= \frac{\text{chord}(1,\theta)}{\text{chord}(r,\theta)} \\ \text{chord}(r,\theta) &= r \cdot \text{chord}(1,\theta) \end{aligned}$$

as desired. □

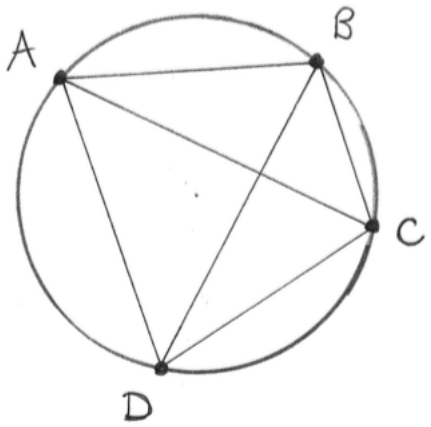
The pure geometers weren't able to make much headway in the computation of chord length, except for a few special angles such as π, $\pi/2$, $\pi/3$, $\pi/4$, $\pi/5$ and $\pi/6$.[52] However, the astronomers didn't have the luxury of waiting for rigorous theorems. Indeed, astronomers must do all of their computations in terms of angles because straight line distances in the heavens are inaccessible to us. The astronomers therefore had to perform many trigonometric computations in their work and in the absence of exact formulas they were happy to accept approximate values.

[52]The angle $\pi/7$ was too difficult.

The most famous astronomer of the ancient world was Claudius Ptolemy (c. 100–170 AD), who lived in Alexandria approximately 450 years after Euclid. In Chapter 1 we discussed Ptolemy's work the *Almagest* which provided the first accurate quantitative model of the universe. In order to obtain this accuracy, Ptolemy needed to be able to compute the approximate chord length for an arbitrary angle, and so Book I of his XIII volume work is devoted to compiling a table of chord lengths. To be specific, Ptolemy fixed the radius $r = 60$[53] and then he computed an approximate value of chord$(60, \theta)$ for each angle θ between 0° and 180°, measured in increments of $(1/2)^\circ$.

The commentator Theon of Alexandria (c. 335–405 AD) tells us that earlier tables of chords were compiled by Hipparchus and Menelaus, but Ptolemy's table of chords is the earliest example that survives. The table itself appears in Book I Chapter 11 of the *Almagest* and Chapter 10 presents the geometric theorems that were used in preparation of the table. Most of the ideas in Chapter 10 come from the *Elements* but there is one remarkably beautiful theorem that is original to the *Almagest* and has become known as "Ptolemy's Theorem".

Ptolemy's Theorem. Consider any four points A, B, C, D on the boundary of a circle:



Then the six distances between these points are related by the equation

$$AC \cdot BD = AB \cdot CD + AD \cdot BD.$$

//

The proof of the theorem is not important (I'm sure you could come up with a proof if you tried hard enough). The point is that Ptolemy's Theorem provides a simple algebraic relationship

[53] We saw above that the radius is not mathematically important; presumably he chose this radius for computational convenience. The fact that Ptolemy used a base-60 system of angle measurement comes from the Babylonian tradition. The ancient Babylonians were talented astronomers who had compiled extensive tables of observations. The Babylonian tables were translated to Greek after the conquest of Alexander the Great in 331 BC and had a strong influence on the Hellenistic astronomers of Alexandria.

between various chord lengths in a circle. In particular, Ptolemy used his theorem to derive formulas relating the chord lengths when angles are added and subtracted.

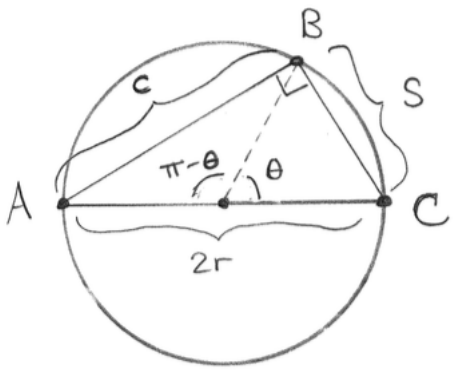
Ptolemy's Angle Sum and Difference Formulas. Consider angles $\theta_1 > \theta_2$ in a circle of radius r and define the following chord lengths:

$$\begin{aligned} s_1 &:= \text{chord}(r, \theta_1) \\ s_2 &:= \text{chord}(r, \theta_2) \\ s_{1+2} &:= \text{chord}(r, \theta_1 + \theta_2) \\ s_{1-2} &:= \text{chord}(r, \theta_1 - \theta_2). \end{aligned}$$

Then we have the following formulas expressing s_{1+2} and s_{1-2} in terms of s_1 and s_2:

$$\begin{aligned} s_{1+2} &= \left(s_1 \cdot \sqrt{4r^2 - s_2^2} + s_2 \cdot \sqrt{4r^2 - s_1^2} \right) / 2r \\ s_{1-2} &= \left(s_1 \cdot \sqrt{4r^2 - s_2^2} - s_2 \cdot \sqrt{4r^2 - s_1^2} \right), 2r \end{aligned}$$

Proof: First let us recall Thales' Theorem from Chapter 1. Consider a triangle ABC inscribed in a circle of radius r. If the segment AC is a diameter of the circle then it follows that $\angle ABC$ is a right angle:



Furthermore, if we denote the side lengths by $s := BC$ and $c := AB$ then it follows from the Pythagorean Theorem that

$$s^2 + c^2 = (2r)^2 = 4r^2.$$

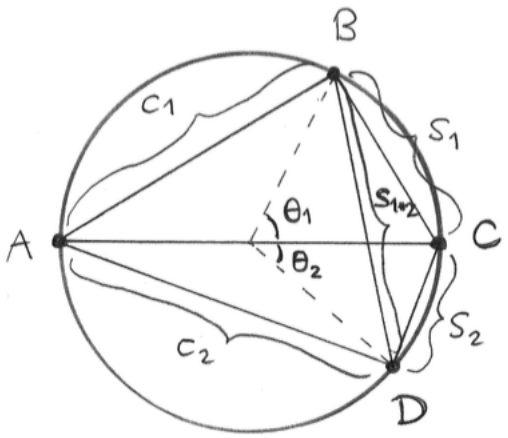
Observe that if $s = \text{chord}(r, \theta)$ is the chord length corresponding to an angle θ then we can think of $c = \text{chord}(r, \pi - \theta)$ as the chord length of the complementary angle $\pi - \theta$.[54]

[54] I have chosen the letters "s" and "c" to stand for "sine" and "cosine". See the Epilogue for more details.

Now let us use Ptolemy's Theorem to prove the angle sum and difference formulas. In the following computations we will denote the complementary chord lengths by

$$c_1 := \text{chord}(r, \pi - \theta_1) = \sqrt{4r^2 - s_1^2},$$
$$c_2 := \text{chord}(r, \pi - \theta_1) = \sqrt{4r^2 - s_2^2}.$$

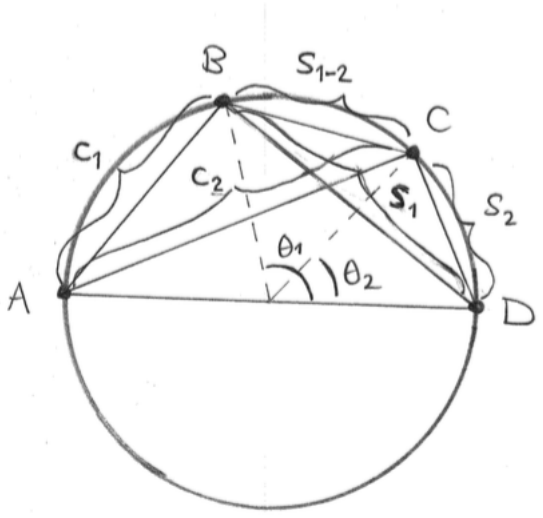
To compute $s_{1+2} = \text{chord}(r, \theta_1 + \theta_2)$ we consider the following inscribed quadrilateral in which the segment AC is a diameter:



Then by applying Ptolemy's Theorem we obtain

$$AC \cdot BD = AB \cdot CD + AD \cdot BD$$
$$(2r) \cdot s_{1+2} = s_2 \cdot c_1 + s_1 \cdot c_2$$
$$(2r) \cdot s_{1+2} = s_2 \cdot \sqrt{4r^2 - s_1^2} + s_1\sqrt{4r^2 - s_2^2}$$

as desired. Finally, to compute $s_{1-2} = \text{chord}(r, \theta_1 - \theta_2)$ we consider the following inscribed quadrilateral in which the segment AD is a diameter:



Then by applying Ptolemy's Theorem we obtain

$$\begin{aligned} AC \cdot BD &= AB \cdot CD + AD \cdot BD \\ s_1 \cdot c_2 &= s_2 \cdot c_1 + (2r) \cdot s_{1-2} \\ s_1 \cdot c_2 - s_2 \cdot c_1 &= (2r) \cdot s_{1-2} \\ s_1 \cdot \sqrt{4r^2 - s_2^2} - s_2 \cdot \sqrt{4r^2 - s_1^2} &= (2r) \cdot s_{1-2} \end{aligned}$$

as desired. □

Recall from Section 3.1 that the Babylonians had an efficient computational method for approximating square roots. Thus, by starting from the chord lengths of a few special angles, Ptolemy was able to use these general rules to compute the approximate length of $\text{chord}(r, \theta)$ for each half-degree angle between 0° and 180°. //

Nevertheless, **exact formulas** for the chord length remained difficult to find. To illustrate the difficulty of this problem, let us return to the classical question of constructing regular polygons.

Question: For which numbers n can a regular n-gon be constructed with straightedge and compass?

Suppose that we are given a circle of radius r. Then the side length of an inscribed n-gon is just the chord corresponding to the angle $2\pi/n$, which is $1/n$-th of the full circle. For convenience we will denote this chord length by

$$\boxed{s_n := \text{chord}(r, 2\pi/n) = r \cdot \text{chord}(1, 2\pi/n).}$$

Now it is clear that the regular n-gon is constructible if and only if we can construct a line segment of length s_n. The first few values are easy to compute:

$$\begin{aligned} s_1 &= 0 \\ s_2 &= 2 \cdot r \\ s_3 &= \sqrt{3} \cdot r \\ s_4 &= \sqrt{2} \cdot r \\ s_6 &= r \end{aligned}$$

If r is constructible then we know that $\sqrt{2} \cdot r$ is constructible because it is just the diagonal of a square with side length r. Furthermore, it is not too difficult to see that $\sqrt{3}$ is constructible. In general, we have the following theorem.

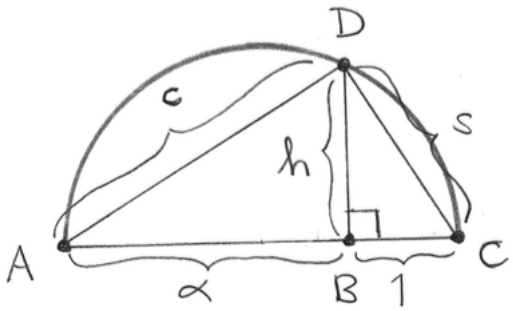
Constructible Line Segments. Suppose we are given line segments of length α and β. Then it is possible to construct line segments of length

$$\alpha + \beta, \quad \alpha - \beta \ \ (\text{if } \alpha > \beta), \quad \alpha \cdot \beta \quad \text{and} \quad \alpha/\beta$$

using only a straightedge and compass. Furthermore, if we are given a line segment of length α then we can construct a line segment of length $\sqrt{\alpha}$. //

Proof: The fact that $\alpha + \beta$ and $\alpha - \beta$ are constructible is immediate and the fact that $\alpha \cdot \beta$ and α/β are constructible is not so interesting. So I'll just prove that $\sqrt{\alpha}$ is constructible.

To do this, we begin with a line segment AB of length α and then we extend the line segment AB to C so that BC has length 1. Now we construct the circle on the diameter AC and extend a perpendicular line from B to D as in the following figure:



I didn't spell out the details, but you will be happy to believe that all of this was possible using compass and straightedge. Finally, I claim that the line segment BD (which we constructed) has length $\sqrt{\alpha}$. To see this, we define the lengths $c := AD$, $h := BD$ and $s := CD$ as in the diagram. On the one hand, the Pythagorean Theorem applied to the right triangles ABD and BCD gives

$$c^2 = \alpha^2 + h^2 \qquad \text{and} \qquad s^2 = 1^2 + h^2.$$

On the other hand, Thales' Theorem tells us that $\angle ADC$ is a right angle. Thus we can apply the Pythagorean Theorem to the right triangle ADC to obtain

$$\begin{aligned}(\alpha+1)^2 &= c^2+s^2\\ \alpha^2+2\alpha+1 &= c^2+s^2\\ \cancel{\alpha^2}+2\alpha+\cancel{1} &= (\cancel{\alpha^2}+h^2)+(\cancel{1}+h^2)\\ 2\alpha &= 2h^2\\ \alpha &= h^2\\ \sqrt{\alpha} &= h\end{aligned}$$

as desired. □

We also have the following general result.

Theorem (Doubling a Regular Polygon). If the regular n-gon is constructible then the regular $2n$-gon is constructible. //

Proof: In Chapter 2 we discussed how to perform this construction by bisecting each angle. Now I'll give an algebraic proof that the construction exists without saying how to do it.

We assume that the radius r and the line segment of length s_n are constructible. Then from Ptolemy's Angle Sum Formula with $s_1 = s_2 = s_n$ and $s_{1+2} = s_{2n}$ we obtain

$$s_{2n} = \left(s_n \cdot \sqrt{4r^2 - s_n^2} + s_n \cdot \sqrt{4r^2 - s_n^2}\right) / 2r = \left(s_n \cdot \sqrt{4r^2 - s_n^2}\right) / r.$$

Since the operations of addition/subtraction, mutiplication/division and square root extraction are constructible this formula shows that a line segment of length s_{2n} (and hence a regular $2n$-gon) is constructible. □

Now assume that we have a circle with constructible radius r. Since whole numbers and square roots are constructible we conclude that the lengths $s_3 = \sqrt{3} \cdot r$, $s_4 = \sqrt{3} \cdot r$ and $s_6 = r$ are constructible and hence we can construct the regular triangle, square and hexagon inscribed in our circle. Furthermore, since any regular polygon can be doubled we conclude that the regular 2^k-gon and $2^k \cdot 3$-gon are constructible for any k.

At this point, we have the following knowledge of constructible polygons:

n	regular n-gon is constructible
3	yes
4	yes
5	?
6	yes
7	?
8	yes
9	?
10	?

You might object that we already knew explicit constructions for these polygons, so the algebraic results above have gained us exactly nothing. That's true. But we won't be able to go any further without algebra. The last big result of the ancient geometers on contructibility of polygons was to show that **the regular pentagon is constructible**. Euclid's Proposition IV.11 gives an explicit and tedious construction. However, an easier way to prove the **existence** of a construction comes from one of the final propositions of the *Elements*, which gives an implicit formula for the side length of a regular pentagon.

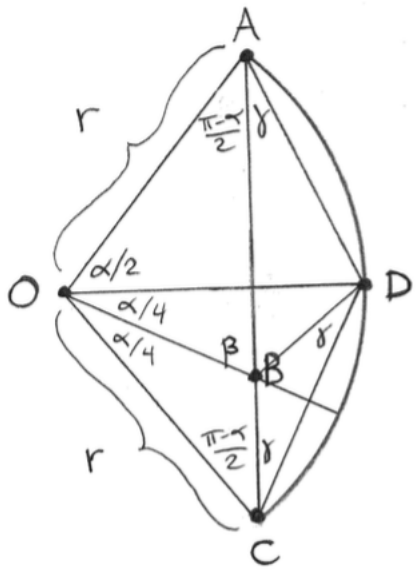
Proposition XIII.10 (Side Length of a Regular Pentagon). *If an equilateral pentagon is inscribed in a circle, then the square on the side of the pentagon equals the sum of the squares on the sides of the hexagon and the decagon inscribed in the same circle.*

In our notation, this says that

$$s_5^2 = s_{10}^2 + s_6^2.$$

//

Proof: Since $s_6 = r$ we want to show that $s_5^2 = s_{10}^2 + r^2$. So consider a circle of radius r and cut out a segment with angle $\alpha = 2\pi/5$. By dividing α into four equal angles we obtain the following diagram (where the angle α is exaggerated to make the diagram more readable):



Note that we have $\angle OAC = \angle OCA = (\pi - \alpha)/2$ because the triangle $\triangle OAC$ is isoceles. For the same reason we know that the three angles labeled by γ are equal to each other. Our first task is to compute the angle $\beta = \angle OBA$. On the one hand, since the angles in the triangle $\triangle OAB$ sum to π we have

$$\beta = \pi - \frac{\pi - \alpha}{2} - \frac{3\alpha}{4} = \frac{2\pi - \alpha}{4}.$$

On the other hand, since $\alpha = 2\pi/5$ we have

$$\begin{aligned} 5\alpha &= 2\pi \\ 4\alpha &= 2\pi - \alpha \\ \alpha &= \frac{2\pi - \alpha}{4} \end{aligned}$$

and it follows that $\alpha = \beta$ (even though it doesn't look that way in my exaggerated diagram). We conlude that the triangles $\triangle OAB$ and $\triangle CAO$ are similar (they have the same angles) and hence

$$\begin{aligned} CA/AO &= OA/AB \\ s_5/r &= r/AB \\ AB \cdot s_5 &= r^2. \end{aligned} \tag{1}$$

Next observe that the triangles $\triangle BCD$ and $\triangle DAC$ are similar because they are both isoceles with base angle γ and it follows that

$$AC/CD = DC/CB$$

$$s_5/s_{10} = s_{10}/CB$$
$$CB \cdot s_5 = s_{10}^2.$$

Finally, by adding equations (1) and (2) we obtain

$$\begin{aligned} AB \cdot s_5 + BC \cdot s_5 &= r^2 + s_{10}^2 \\ (AB + BC) \cdot s_5 &= r^2 + s_{10}^2 \\ s_5 \cdot s_5 &= r^2 + s_{10}^2 \\ s_5^2 &= s_{10}^2 + r^2 \end{aligned}$$

as desired. □

[Remark: It follows from this result and the Pythagorean Theorem that a triangle with side lengths s_{10}, s_6 and s_5 has a right angle between the sides s_{10} and s_6. It is interesting that no such triangle appears in the proof.]

For Euclid it was preferable to state this result in geometric terms but we moderns would like to have an algebraic formula for s_5. We can do this with a bit of High School algebra.

In order to make the notation cleaner we will assume that $r = 1$. First note that we obtain a second equation relating s_5 and s_{10} from Ptolemy's Angle Sum Formula. By setting $\theta_1 = \theta_2 = 2\pi/10$ so that $\theta_1 + \theta_2 = 2\pi/5$ we obtain

$$\begin{aligned} s_5 &= s_{10} \cdot \sqrt{4 - s_{10}^2} \\ s_5^2 &= \left(s_{10} \cdot \sqrt{4 - s_{10}^2} \right)^2 \\ s_5^2 &= s_{10}^2 \left(4 - s_{10}^2 \right). \end{aligned}$$

Then we substitute this expression together with $s_6 = r = 1$ into Euclid's equation to obtain

$$\begin{aligned} s_5^2 &= s_{10}^2 + s_6^2 \\ s_{10}^2 \left(4 - s_{10}^2 \right) &= s_{10}^2 + 1 \\ 4 \cdot s_{10}^2 - (s_{10}^2)^2 &= s_{10}^2 + 1 \\ 0 &= 1 \cdot (s_{10}^2)^2 - 3 \cdot s_{10}^2 + 1. \end{aligned} \tag{3}$$

We have arrived at a "quadratic equation" for the quantity s_{10}^2. The Greeks would have regarded this as fairly meaningless because it does not have any obvious geometric interpretion. The modern theory of equations was developed by the mathematician, astronomer and geographer Al-Khwarizmi (c. 780–850 AD) who worked at the House of Wisdom in Baghdad. In fact, we take the word "algebra" (Arabic: *al-jabr*) from the title of his book: *al-Kitab al-mukhtasar fi hisab al-jabr wal-muqabala* (c. 830 AD) (The Compendious Book on Calculation by Completion and Balancing). The point of Al-Khwarizmi's theory is that we should

temporarily **ignore** any geometric meaning contained in the equation (3) and simply proceed by a mechanical process of computation.[55] Today we know this mechanical process as the *quadratic formula*, and in the case of equation (3) it tells us that

$$s_{10}^2 = \frac{-(-3) \pm \sqrt{(-3)^2 - 4(1)(1)}}{2(1)} = \frac{3 \pm \sqrt{5}}{2}.$$

Since the quantity s_{10}^2 is positive we throw away the negative value of the square root to obtain $s_{10}^2 = (3 + \sqrt{5})/2$ and then we substitute this value back into Euclid's equation to obtain

$$\begin{aligned} s_5^2 &= s_{10}^2 + s_6^2 \\ s_5^2 &= (3 + \sqrt{5})/2 + 1 \\ s_5^2 &= (3 + \sqrt{5})/2 + 2/2 \\ s_5^2 &= (5 + \sqrt{5})/2 \\ s_5 &= \sqrt{(5 + \sqrt{5})/2} \\ \text{chord}(1, 2\pi/5) &= \sqrt{(5 + \sqrt{5})/2}. \end{aligned}$$

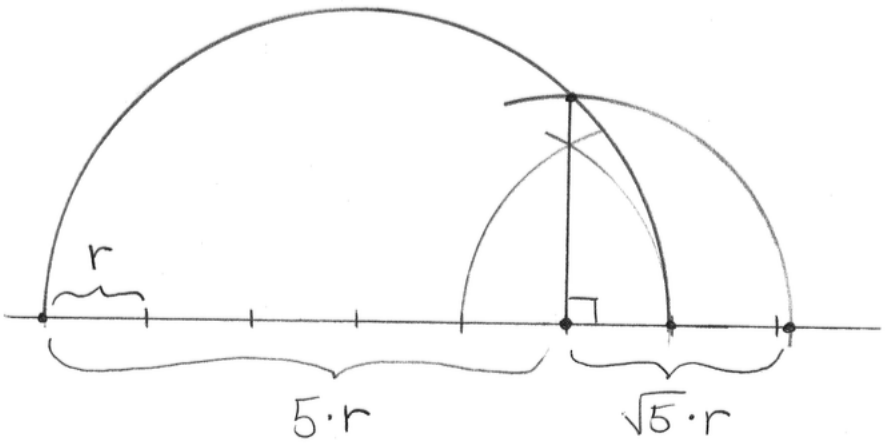
Finally, we obtain the chord length for an arbitrary radius r:

$$s_5 = \text{chord}(r, 2\pi/5) = r \cdot \text{chord}(1, 2\pi/5) = r \cdot \sqrt{\frac{5 + \sqrt{5}}{2}}.$$

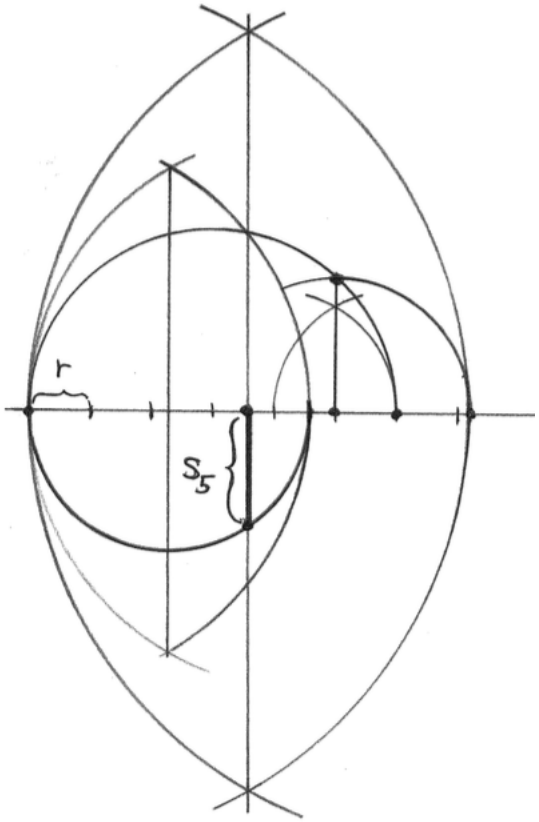
Now we see the value of algebra for the study of constructible polygons: The number $\sqrt{(5 + \sqrt{5})/2}$ is complicated, but since it is formed from the constructible numbers 2 and 5 via the constructible operations of addition/subtraction, addition/multiplication and the extraction of square roots, we conclude that **the regular pentagon is constructible with straightedge and compass**.

In fact, this formula gives us a recipe for the construction. First we use our square root trick to construct a line segment of length $r \cdot (5 + \sqrt{5})$:

[55] It is no surprise that the word "algorithm" is based on the Latin version of Al-Khwarizmi's name.



Then we bisect this segment, add r, and bisect it again. Finally, we do the square root trick a second time to obtain a line segment of length $s_5 = r \cdot \sqrt{(5+\sqrt{5})/2}$:



At this point it's easy to construct the regular pentagon but I won't bother because nobody really cares about constructing regular pentagons. I just wanted to illustrate to you that the

process is highly nontrivial. //

We have seen that the regular n-gon is constructible when $n = 2^k$, $n = 2^k \cdot 3$ or $n = 2^k \cdot 5$ for any k. And this is where human knowledge stood for over two thousand years until 1796 when the young Carl Friedrich Gauss[56] proved the shocking result that

the regular 17*-gon is contructible!*

As you can imagine, he did not actually provide a construction (because no one really cares about constructing the regular 17-gon) but instead he showed that the side length s_{17} has a formula that can be computed in terms of the constructible operations of addition/subtraction, multiplication/division and extracting square roots. The explicit formula is so long that it doesn't even fit on one line:

$$s_{17} = \frac{r}{4} \cdot \sqrt{2} \cdot \sqrt{17 - \sqrt{17} - \sqrt{2} \cdot \left(\sqrt{\alpha} + \sqrt{17 - \sqrt{17}} \right)}, \quad \text{where}$$

$$\alpha = 34 + 6 \cdot \sqrt{17} + \sqrt{2} \cdot \left(\sqrt{17} - 1 \right) \cdot \sqrt{17 - \sqrt{17}} - 8 \cdot \sqrt{2} \cdot \sqrt{17 + \sqrt{17}}.$$

After seeing the compass-and-straightedge construction of s_5 you can imagine that the construction of s_{17}, while logically possible, is not practical for a human geometer.[57]

But why 17 ? Why didn't Gauss find a construction for the regular 7-gon or 9-gon, which were both open problems at the time? The reason is that the 7-gon and 9-gon are **impossible to construct with compass and straightedge**. After thinking about the problem for five years, Gauss stated the following result without proof in his *Disquisitiones Arithmeticae* (1801). A full proof was supplied by Pierre Wantzel in 1837.

The Gauss-Wantzel Theorem (1796–1837). For any positive integer n, let $\varphi(n)$ count the numbers between 1 and n that share no common factors with n.[58] Then the regular n-gon is constructible with straightedge and compass if and only if $\varphi(n)$ is a power of 2. Moreover, if $\varphi(n) = 2^k$ then an algebraic expression for s_n will require k levels of nested square roots. //

This is a result of abstract algebra and number theory, so unfortunately I cannot explain it in this course. But let's add the quantity $\varphi(n)$ to our table to see if it agrees with our previous results:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\varphi(n)$	1	1	2	2	4	2	6	4	6	4	10	4	12	6	8	8	16
constructible?	Y	Y	Y	Y	Y	Y	N	Y	N	Y	N	Y	N	N	Y	Y	Y

[56] him again!

[57] Wikipedia probably has a computer animation of the construction.

[58] The fancy name for this is "Euler's totient function". It was given this name by a strange mathematician named James Joseph Sylvester.

We see that the regular n-gon is **not constructible** when $n = 7, 9, 11, 13, 14$ because in these cases $\varphi(n)$ is not a power of 2. For example, here are the numbers from 1 to 14 with the numbers that share a common factor with 14 crossed out:

$$1\ \cancel{2}\ 3\ \cancel{4}\ 5\ \cancel{6}\ \cancel{7}\ \cancel{8}\ 9\ \cancel{10}\ 11\ \cancel{12}\ 13\ \cancel{14}$$

Since there are six numbers not crossed out we verify that $\varphi(14) = 6$. Then since 6 is not a power of 2 the Gauss-Wantzel theorem tells us that the regular 14-gon cannot be constructed with straightedge and compass. Thus the question of constructibility of regular polygons has been transformed into the following question of number theory.

Question: For which values of n is the totient function $\varphi(n)$ a power of 2?

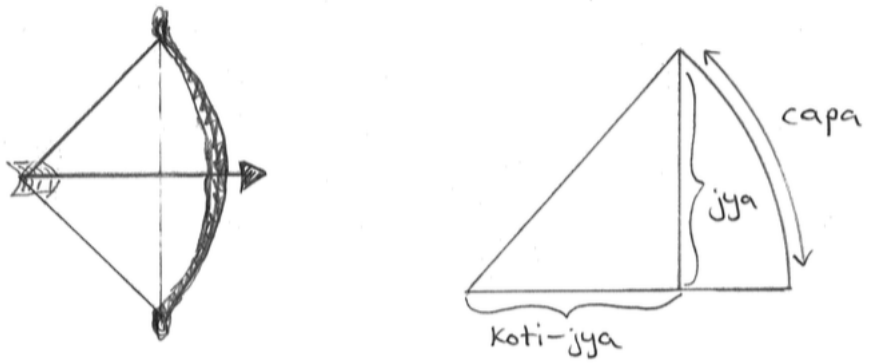
One can show that $\varphi(n)$ is a power of 2 precisely when n equals a power of 2 multiplied by a set of distinct prime numbers of the form $2^{2^p} + 1$ (called *Fermat prime numbers*). Unfortunately the Fermat prime numbers are not completely understood[59] so in some sense the problem is still open.

Epilogue: From Jya-Ardha to Sine. Throughout this section I have avoided the modern notation for trigonometry because it is rather terrible and it tends to obscure the motivating problems behind the subject. As with our modern symbol π, the modern language for trigonometry in terms of "sin, cos, tan" and all that nonsense was standardized by Leonhard Euler's textbook *Introductio in analysin infinitorum* (1748) (Introduction to Analysis of the Infinite). Where did the seemingly random names "sin, cos, tan" come from?

I mentioned above that the most sophisticated trigonometry of the Hellenistic world was done by Claudius Ptolemy (c. 100–170 AD) in the *Almagest*. The next burst of progress came during the Gupta Empire of India. While earlier work in trigonometry had focused on the chord function, the mathematician-astronomer Aryabhata (c. 476–550 AD) provided a table of "half-chords" in his work the *Aryabhatiya* (499 AD).

The Indian notation for trigonometry was quite natural. They noticed that the chord of a circle subtended by a given angle looks like a pulled bowstring:

[59] For example, we don't know if there are infinitely many of them.

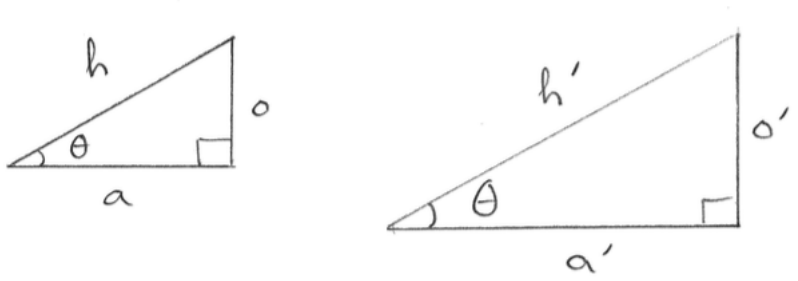


For this reason they referred to the chord and the circular arc by the Sanskrit words *jya* and *capa* which mean "bowstring" and "bow", respectively. They used the word *koti-jya* for the distance between the center of the circle and the midpoint of the chord (*koti* means "point" or "cusp".) Then at some point, for whatever reason, they began to refer to the *jya-ardha* (literally "half-bowstring") instead of the *jya*. Aryabhata's work contained a table of *jya-ardha* for various angles and for the fixed radius $r = 3438$ (which is the approximate number of arc-minutes contained in one radian).

The next stage of development occurred during the Islamic Golden Age in Baghdad. The Islamic mathematicians and astronomers had access to Arabic translations of both the Hellenistic and Indian works of astronomy, thus they were able to compare them and take the best from both. At first the Sanskrit term *jya* was transliterated into the nonsense term *jiba* or *jyb*. However, since vowels are often removed from the Arabic script, the nonsense word *jiba/jyb* was later mistaken for the Arabic word *jaib*, which means "bosom", "fold of a garment", or "bay". Meanwhile, Ptolemy's work had been lost in Europe and was recovered in Arabic translation. When Gerard of Cremona (c. 1114–1187) translated the Arabic version of Ptolemy's *Almagest*[60] into Latin he interpreted the word *jaib* as *sinus*, which has similar meanings, and this became our modern *sine* function.

To see the relationship, let me recall the modern school definitions of sin, cos and tan. Consider the following right triangles with side lengths labeled a, o, h, a', o', h' (for "adjacent", "opposite" and "hypotenuse"):

[60] In fact, the title *Almagest* comes from this translation. The original Greek title was *Mathematike Syntaxis* (Mathematical Treatise) which later became *Megale Syntaxis* (The Great Treatise) or *Megiste Syntaxis* (The Greatest Treatise). This was translated into Arabic as *al-majisti* and finally into Latin as *Almagest*.



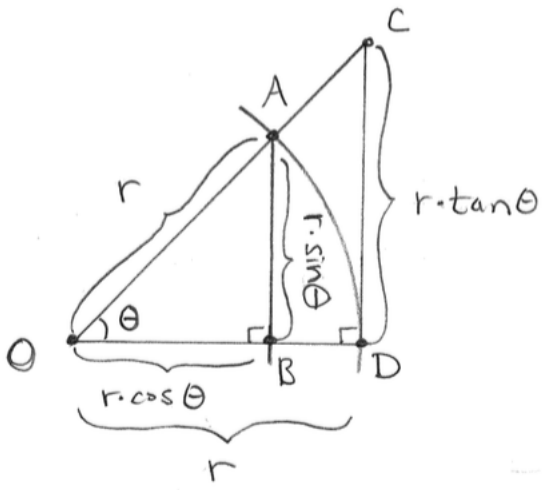
If the angle θ is common to both triangles then the third angle must also be equal (because the angles in each triangle sum to π). Therefore by Euclid's Proposition VI.4 (proportionality of similar triangles) the following ratios are well-defined:

$$\sin\theta := \frac{o}{h} = \frac{o'}{h'}, \quad \cos\theta := \frac{a}{h} = \frac{a'}{h'} \quad \text{and} \quad \tan\theta := \frac{o}{a} = \frac{o'}{a'}.$$

Now let $h = r$ be the radius of a circle so that $o = r \cdot \sin\theta$ is the *jya-ardha* (or half-chord) and $a = r \cdot \cos\theta$ is the *koti-jya*. The "jya" explains the "sine" and the "koti" explains the "co".[61] Then by considering a right triangle with angle $\theta/2$ we obtain the following expression for the full chord length in terms of the sine function:

$$\boxed{\text{chord}(r, \theta) = 2r \cdot \sin(\theta/2).}$$

And what about "tan"? To understand this we need to look back to Archimedes' proof for the area of a circle. By considering the inscribed and circumscribed 2^n-gons for a circle of radius r we obtained the following diagram with $\theta = \pi/2^n$:



[61] Or maybe the "co" refers to the "complementary" angle $\pi - \theta$ between sides h and o.

This diagram explains the meaning of "tan": it stands for "tangent" because the line segment of length $r \cdot \tan\theta$ in the diagram is tangent to the circle. This also explains why I used the letters "s" and "t" in Archimedes' proof (i.e., to stand for "sin" and "tan"). I didn't use "c" at that time because it would have been confused with "circumference".

Finally, recall that the purpose of this diagram in Archimedes' proof was to compare the areas of the inscribed and the circumscribed polygon to the area of the circle. Now that we know the formula πr^2 for the area of the circle we can reinterpret this argument in an interesting way. If we measure the angle θ in radians then we observe that the sector of the circle (the pizza slice) defined by θ contains $\theta/2\pi$ of the full area of the circle. In other words:

$$(\text{area of sector OAD}) = \frac{\theta}{2\pi} \cdot (\text{area of circle}) = \frac{\theta}{2\pi} \cdot \pi r^2 = \frac{r^2\theta}{2}.$$

But this sector is contained between the two right triangles in the diagram, so that

$$\begin{array}{ccccc} (\text{area of triangle OAB}) & < & (\text{area of sector OAD}) & < & (\text{area of triangle OCD}) \\ (r\sin\theta)(r\cos\theta)/2 & < & (r^2\theta)/2 & < & r \cdot (r\tan\theta)/2 \\ \sin\theta\cos\theta & < & \theta & < & \tan\theta. \end{array}$$

If we rewrite $\tan\theta$ as $\sin\theta/\cos\theta$ and then divide all three expressions by $\sin\theta$ (which is a **positive** number because $0 < \theta < \pi$) then we obtain the inequalities

$$\begin{array}{ccccc} \sin\theta\cos\theta & < & \theta & < & \sin\theta/\cos\theta \\ \cos\theta & < & \theta/\sin\theta & < & 1/\cos\theta. \end{array}$$

So what? Well, if the angle θ is very small (i.e., close to zero) then the quantities $\cos\theta$ and $1/\cos\theta$ are both very close to 1. Since the ratio $(\theta/\sin\theta)$ is **squeezed** between two quantities that both approach 1, we conclude that

the ratio $\frac{\theta}{\sin\theta}$ *approaches* 1 *as the angle* θ *approaches zero.*

It turns out that this harmless statement, which was foreshadowed in Archimedes' computation of circular area, is the key fact upon which all of modern trigonometry is based.

The mathematician and astronomer Madhava of Sangamagrama (c. 1340–1425) was the founder of a thriving mathematical school in the Kerala region of India. By using the ratio $(\theta/\sin\theta)$ he was able to find explicit (but infinite) formulas for the basic trigonometric functions. These ideas were rediscovered later by Isaac Newton and Gottfried Leiniz in the 1670s as part of their discovery of the Calculus. The modern algebraic formula for the sine function is

$$\sin\theta = \theta - \frac{\theta^3}{3\cdot 2\cdot 1} + \frac{\theta^5}{5\cdot 4\cdot 3\cdot 2\cdot 1} - \frac{\theta^7}{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1} + \cdots.$$

For this formula to be valid we need to measure θ in radians and we need to carry out the computation to infinity. If we truncate the computation after several steps then we will obtain a good approximate value for $\sin\theta$.[62]

[62] This is how your calculator does it.

In retrospect, we realize that the sine function (and hence the chord function) is *transcendental*, meaning that it is impossible to compute with a finite number of algebraic operations. This is the ultimate reason why "trigonometry is hard" and why the subject is a lot more modern than one would expect.

3.5 Rigorous and Intuitive Mathematics

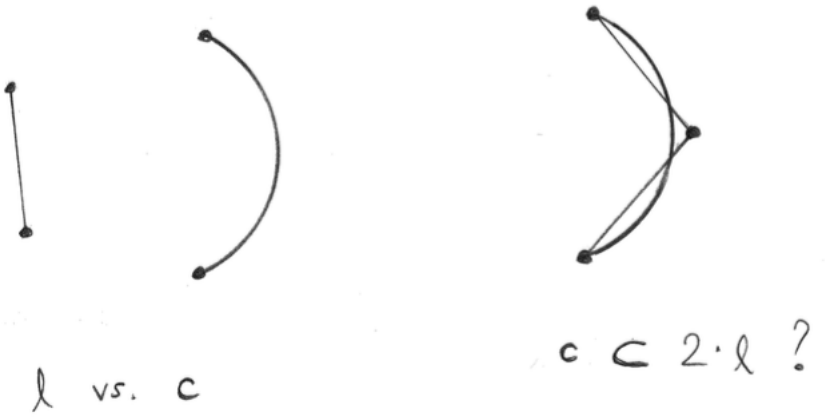
In section 3.3 we saw Archimedes' proof from *Measurement of a Circle* that the area of a circle is given by

$$A = \frac{1}{2}Cr,$$

where C is the circumference and r is the radius. Since Euclid's *Elements* provides no way to talk about the length of a curved path, Archimedes needed to introduce some new axioms in order to make the proof rigorous. He introduced these axioms (postulates) explicitly in an accompanying two-volume work called *On the Sphere and Cylinder*; perhaps not coincidentally, he had exactly five postulates. Here they are in modern language:

> **Postulate 1.** The shortest path between two points is a straight line.
>
> **Postulate 2.** Given two concave paths on the same side of a straight line, the path closer to the line is shorter.
>
> **Postulate 3.** Among surfaces whose boundary lies in a plane, the flat surface has the least area.
>
> **Postulate 4.** Given two concave surfaces on the same side of a flat plane with their common boundary in the plane, the surface closer to the plane has less area.
>
> **Postulate 5.** Any two geometric magnitudes of the same kind "have a ratio".

We already discussed Postulates 1 and 2 in the previous section. They allow comparisons between lengths of curved paths. Postulates 3 and 4 do the same thing for areas of curved surfaces. Archimedes Postulate 5 is a direct extension of Euclid's Definition V.5 which we discussed in Section 3.2. There Euclid defined what it means for two geometric magnitudes to "have a ratio", i.e., when either can be contained in a multiple of the other. It is clear that any two straight line segments, areas or volumes "have a ratio" in this sense, but is unclear how the definition should apply to curved paths and surfaces. For example, consider the following line segment of length ℓ and curve of length c:



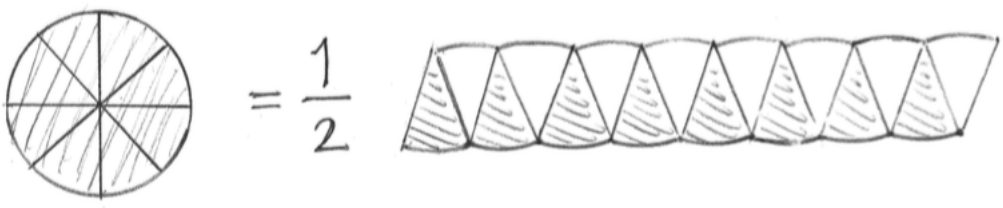
For ℓ and c to "have a ratio" in the Euclidean sense there must exist whole numbers m and n such that $m \cdot \ell \subset c$ and $\ell \subset n \cdot c$. We see on the right of the figure that c and $2 \cdot \ell$ are sort of comparable, but it is clear that c will never fit completely inside $n \cdot \ell$ for any whole number n. To get around this problem, Archimedes simply **declared** that ℓ and c **do** have a ratio so he could go about the business of computing it.

In the previous section we saw that Archimedes' Theorem $A = \frac{1}{2}Cr$ can be proved rigorously from the results in Euclid's *Elements* by adding Postulates 1 and 2 (and I suppose it also needs Postulate 5), but

did you find the proof convincing?

Before answering, let me show you a "non-rigorous" proof that $A = \frac{1}{2}Cr$.

Non-Rigorous Proof. Consider a circle with area, circumference and radius given by A, C and r, respectively. Now cut the circle into $2n$ equal sectors, just like slicing a pizza. By rearranging the pieces of pizza we can form exactly one half of a shape that is approximately a rectangle with side lengths C and r:

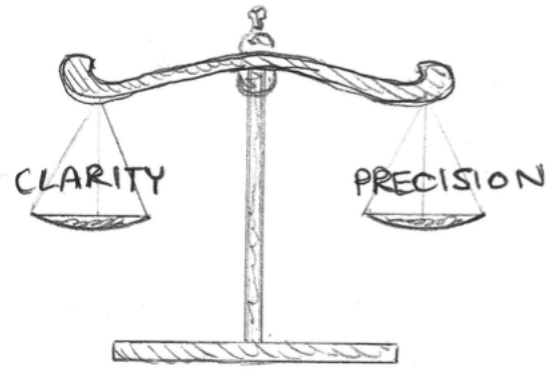


It seems that we can make the shape on the right arbitrarily close to a perfect rectangle by taking n large enough. Thus in the limit we must have $A = \frac{1}{2}Cr$. □

I'll bet you find this argument more convincing than the official (rigorous) proof from *Measurement of a Circle*. This illustrates a strange phenomenon in mathematics:

sometimes a proof gives no hint of how the theorem was discovered.

In practice, every mathematical proof has to make a compromise between rigor (precision) and readability (clarity):



The style of proof in Euclid's *Elements* favors precision over clarity. The goal here is to provide a certificate of absolute truth, not necessarily to persuade an audience. It is the reader's job to do the hard work of deciphering the theorems.

For most of the modern era, mathematicians had access to Archimedes' Euclidean-style proofs for properties of circles and spheres, but they had no hint of how Archimedes had discovered the theorems in the first place. Then in 1906, a momentous discovery was made: a lost work of Archimedes called *The Method of Mechanical Theorems* was found hidden underneath a 13th century Christian religious text written on parchment (i.e., stretched and bleached animal skin). Since parchment was valuable, it was often scraped clean to make way for a new text. In this case the 13th century monks had erased a 10th century Byzantine Greek copy of works by Archimedes. But the erasure was not perfect and scholars were able to recover the text of Archimedes hiding underneath the religious text. The parchment (known as the *Archimedes Palimpsest*[63]) is now on display at the Cambridge University Library.

In *The Method of Mechanical Theorems* (henceforth known as *The Method*), Archimedes describes the intuitive process by which he arrived at many of his geometric theorems. In this section I will present his fundamental results from *On The Sphere and Cylinder*, but instead

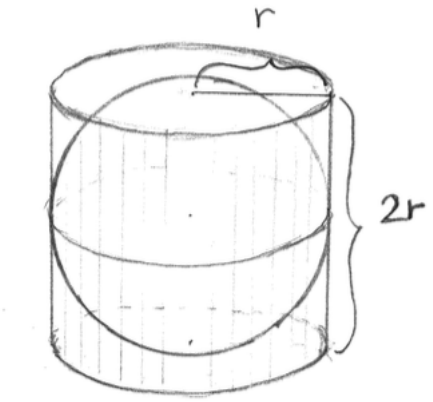
[63] A *palimpsest* is a piece of writing material on which the original has been erased to make room for new writing, but on which traces of the original are still legible.

of following the elaborate Euclidean-style proofs I will follow a process closer to the one that Archimedes describes in *The Method*. That is, my primary goal will be to **convince you**[64] that the results are true.

Here is the big theorem that we will prove

Theorem (Volume and Surface Area of a Sphere). Consider a sphere inscribed in a right cylinder. Then the volume of the sphere is 2/3 of the volume of the cylinder and the surface area of the sphere is equal to the surface area of the side of the cylinder (excluding the top and bottom circles). //

You can see why Archimedes called his work *On the Sphere and Cylinder*. In modern terms we prefer to express the volume and surface area as algebraic formulas in terms of the radius and the universal constant π. Suppose the sphere has radius r as in the following picture:



We already know that the area of the base circle is πr^2 so volume of the cylinder is given by

$$\begin{aligned}(\text{volume of cylinder}) &= (\text{area of base})(\text{height}) \\ &= (\pi r^2)(2r) \\ &= 2\pi r^3.\end{aligned}$$

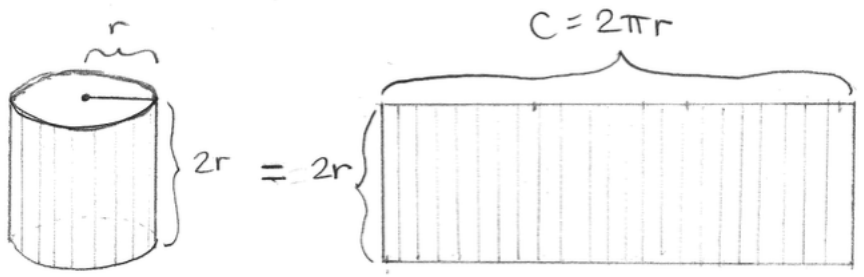
Therefore the theorem states that

$$(\text{volume of sphere}) = \frac{2}{3}(\text{volume of cylinder}) = \frac{2}{3}(2\pi r^3) = \frac{4}{3}\pi r^3.$$

To compute the surface area of the cylinder we can make a vertical cut and unwrap the surface into a rectangle[65] whose height is $2r$ and whose base equals the circumference $C = 2\pi r$:

[64] yes you

[65] Recall from the previous chapter that the surface of a cylinder has zero Gaussian curvature.



Therefore the theorem states that

$$\begin{aligned}(\text{surface area of sphere}) &= (\text{area of side of cylinder}) \\ &= (\text{area of unwrapped rectangle}) \\ &= (\text{base})(\text{height}) \\ &= (2\pi r)(2r) \\ &= 4\pi r^2.\end{aligned}$$

//

So let's get started.

Area of a triangle.

Cavalieri's Principle.

Volume of a parallelopiped.

Volume of a tetrahedron.

Volume of a cone.

Volume of a sphere.

Surface area of a sphere.

3.6 Impossible Problems

Axioms for measurement. Wallace-Bolyai-Gerwien Theorem. Hilbert's third problem. Dehn's Theorem. "Measure theory" is impossible.

Squaring the circle, trisecting angles, doubling cubes.

4 Coordinate Geometry and Transformations

5 Projective Geometry