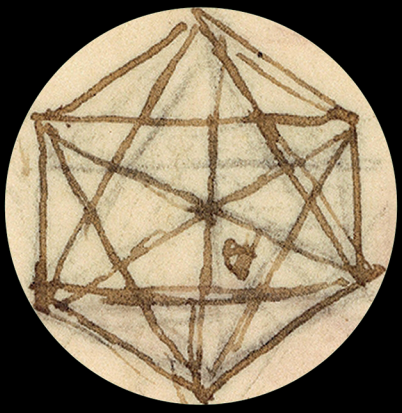
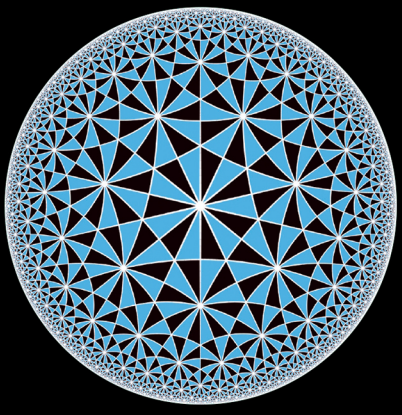
The Story of Proof

Logic and the History of Mathematics

John Stillwell





THE STORY OF PROOF

THE STORY OF PROOF

LOGIC AND THE HISTORY OF MATHEMATICS

John Stillwell

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
99 Banbury Road, Oxford OX2 6JX

press.princeton.edu

ISBN 9780691234366
ISBN (e-book) 9780691234373

Library of Congress Control Number: 2022932472

British Library Cataloging-in-Publication Data is available

Editorial: Diana Gillooly, Kristen Hop, Kiran Pandey
Production: Jacqueline Poirier
Publicity: Matthew Taylor, Carmen Jimenez
Copyeditor: Patricia J. Watson

Jacket images: (*Top*): Leonardo da Vinci (1452–1519), Codex Atlanticus, sheet 518 recto, © Veneranda Biblioteca Ambrosiana / Metis e Mida Informatica / Mondadori Portfolio / Bridgeman Images. (*Left*): Five tetrahedra in a dodecahedron, courtesy of the author. (*Right*): Order-3 heptakis heptagonal tiling, courtesy of Claudio Rocchini.

This book has been composed in MinionPro

Printed on acid-free paper. ∞

Printed in the United States of America

1 3 5 7 9 10 8 6 4 2

To Elaine, my *sine qua non*

Contents

Preface

Proof is the glory of mathematics—and its most characteristic feature—yet proof itself is not considered an interesting topic by many mathematicians. In the United States proof is not deemed an essential part of mathematics education until upper university level, where "introduction to proof" courses are offered. Yet by withholding the concept of proof, we prevent students from seeing how mathematics actually *works*. I thought about naming this book "How Mathematics Works," before deciding on a more modest but accurate title. It is about proof—not just about what proof is but about where it came from, and perhaps where it is going.

We know that mathematics has a logical structure, but we also know that this structure is ever-changing, reflecting its evolution in the collective human mind. There is generally more than one way to prove a given theorem or to develop a given theory. Often the way first discovered is not the simplest or most natural, but vestiges of the old ways survive because of historical inertia or because they appeal to human senses or psychology. For example, geometry continues to appeal to human visual intuition even though it can be done by the symbolic methods of algebra or analysis. Because of this, the mathematical experience is greatly enriched by awareness of historical and logical issues, and we owe it to our students to present mathematics as a rich experience. Even professional mathematicians will be enlightened, I believe, by seeing the evolution of proof in mathematics, because advances in mathematics are often advances in the concept of proof.

A major theme of the book is the relation between logic and computation, where "computation" is understood broadly to include classical algebra. In ancient Greece logic was strong (though deployed mainly in geometry) and computation was weak. In ancient China and India, computation ruled, as it did in Europe when algebra arrived from India via the Muslim world. Then in the seventeenth century Europe made the

further step to *infinitesimal* algebra, namely calculus, which dominated mathematics (and physics) for the next two centuries. Leibniz, in unpublished work, dreamed of reducing logic itself to algebraic calculation. The dream of Leibniz began to take shape when Boole in 1847 created what we now call *Boolean algebra*, thereby reducing a significant fragment of logic to genuine calculation.

But the full logic of mathematics, and the full concept of computation, was not well understood until the twentieth century. In 1879 Frege described a logic adequate for mathematics, but the very idea that logic and computation were mathematical *concepts*, rather than just mathematical methods, did not arise until the 1920s. When this happened, through the work of Post, Gödel, Turing, and others, logic and computation became actual branches of mathematics—indeed, essentially the *same* branch.

Unfortunately, logic and computation developed largely in isolation from the rest of mathematics, so they are less well known than they should be in the mathematical community.[1] This book tries to remedy this situation, by presenting logic as it developed in mainstream mathematics. The history of mathematics can be viewed as a history of proof, because mathematics presents the most extreme challenges to proof: the Pythagorean discovery of irrational numbers, the sixteenth-century encounter with imaginary numbers, the seventeenth-century controversies over infinitesimals, and the nineteenth-century struggles with infinity, to mention just a few.

A second and related theme of the book is the development of concepts, since proofs can often be articulated only when suitable abstract concepts and notation are available to express them. This is seen most clearly in the development of algebra, where many abstractions originated and later spread to other parts of mathematics. But concept development is also a key to geometry and analysis, where concepts that now seem obvious, such as "area" and "limit," emerged only after long struggles with provisional concepts that failed to capture exactly what was intended.

In fact, the network of mathematical concepts is about as complex as the network of theorems, and I have tried to highlight both theorems and concepts by writing them in bold where they make key appearances in

1. Weil (1950) described logic as the "hygiene of the mathematician, it is not his source of food," as though logicians were sanitation workers.

the story. In the early chapters, new concepts are simple and infrequent enough to be defined informally, but later chapters make more formal definitions, particularly when several new concepts arise together and depend on each other.

I hope this book clarifies the role of logic, computation, and abstraction in mathematics for a general mathematical audience and hence gives a better understanding of the nature of proof. It is not an introduction to proof as much as a panoramic view of proof in basic mathematics. Many of the perennial concerns of all mathematicians—such as the relations among geometry, algebra, and analysis and their seemingly different styles of proof—are seen afresh from the viewpoint of logic and history. We see the intuitive origins of concepts, the search for axioms that capture intuition, new intuitions that emerge from axioms, and the connections among geometry, algebra, and analysis that axioms bring to light. It is fairly well known, for example, that Hilbert in the 1890s filled the gaps in Euclid's axioms for geometry. It is less well known that in doing so Hilbert found new connections among geometry, algebra, and even analysis. These connections are explained in chapters 3 and 11.

The arrangement of the book is partly chronological, partly by topic. Fields of mathematics are introduced in chronological order: geometry and number theory, algebra, algebraic geometry, calculus, and so on. But sometimes we follow a particular topic over a long period, so as not to break a train of thought, before turning to the next topic in chronological order. For example, in chapter 4 the story of algebra is told from ancient times until the nineteenth century, because it is mostly self-contained. The influence of algebra on other fields of mathematics, such as geometry, calculus, and number theory, is then told in chapters 5, 6, and 7.

Arranging material by topic also serves to arrange methods of proof, because of the different methods of proof in different fields mentioned above. Today, these methods are so different that practitioners in one field often fail to understand those in another. Among other things, I hope this book will contribute to mutual understanding by explaining the methods of proof characteristic of different fields. It should be accessible to senior undergraduates and also of interest to their teachers—possibly serving as a bridge between my two previous books, *Elements of Mathematics* and *Reverse Mathematics* (Stillwell 2016, 2018).

As always, I thank my wife, Elaine, for her eagle eye in proofreading the manuscript. I also thank Mark Hunacek and the anonymous reviewers for helpful suggestions and corrections.

John Stillwell
South Melbourne, 2021

CHAPTER 1

Before Euclid

The signature theorem of mathematics is surely the **Pythagorean theorem**, which was discovered independently in several cultures long before Euclid made it the first major theorem in his *Elements* (book 1, proposition 47). All the early roads in mathematics led to the Pythagorean theorem, no doubt because it reflects both sides of basic mathematics: number and space, or arithmetic and geometry, or the discrete and the continuous.

The arithmetic side of the Pythagorean theorem was observed in remarkable depth as early as 1800 BCE, when Babylonian mathematicians found many triples $\langle a, b, c \rangle$ of natural numbers such that $a^2 + b^2 = c^2$. Whether they viewed each triple a, b, c as sides of a right-angled triangle has been questioned; however, the connection was not missed in ancient India and China, where there were also geometric demonstrations of particular cases of the theorem.

Nevertheless, the Pythagoreans are rightly associated with the theorem because of their discovery that $\sqrt{2}$, the hypotenuse of the triangle with unit sides, is **irrational**. This discovery was a turning point in Greek mathematics, even a "crisis of foundations," because it forced a reckoning with *infinity* and, with it, the need for *proof*. In India and China, where irrationality was overlooked, there was no "crisis," hence no perceived need to develop mathematics in a deductive manner from self-evident axioms.

The nature of irrational numbers, as we will see, is a deep problem that has stimulated mathematicians for millennia. Even in antiquity, with Eudoxus's theory of proportions, the Greeks took the first step from the discrete toward the continuous.

1.1 THE PYTHAGOREAN THEOREM

For many people, the Pythagorean theorem is where geometry begins, and it is where proof begins too. Figure 1.1 shows the pure geometric form of the theorem: for a right-angled triangle (white), the square on the hypotenuse (gray) is equal to the sum of the squares on the other two sides (black).

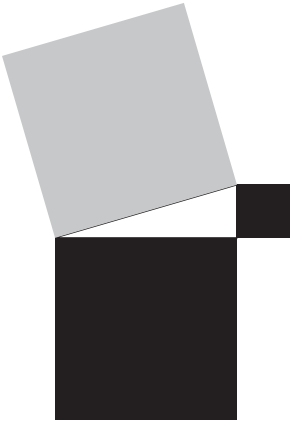


Figure 1.1 : The Pythagorean theorem

What "equality" and "sum" mean in this context can be explained immediately with the help of figure 1.2. Each half of the picture shows a large square with four copies of the triangle inside it. On the left, the large square minus the four triangles is identical with the square on the hypotenuse. On the right, the large square minus four triangles is identical with the squares on the other two sides. Therefore, the square on the hypotenuse *equals* the sum of the squares on the other two sides.

Thus we are implicitly assuming some "common notions," as Euclid called them:

1. Identical figures are equal.
2. Things equal to the same thing are equal to each other.
3. If equals are added to equals the sums are equal.
4. If equals are subtracted from equals the differences are equal.

These assumptions sound a little like algebra, and they are obviously true for numbers, but here they are being applied to geometric objects.

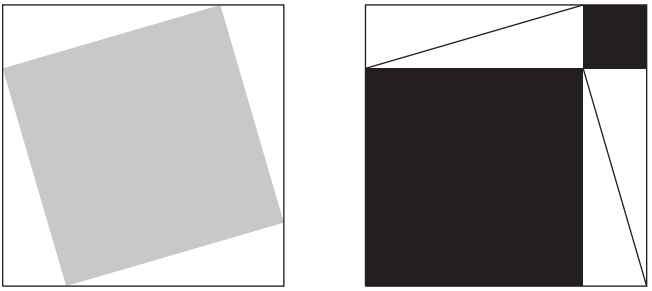


Figure 1.2 : Seeing the Pythagorean theorem

In that sense we have a purely geometric proof of a geometric theorem. The reasons why the Pythagoreans wanted to keep geometry pure will emerge in section 1.3 below.

Although figure 1.2 is as convincing as a picture can be, some might quibble that we have not really explained why the gray and black regions are squares. The Greeks who came after Pythagoras did indeed quibble about details like this, due to concerns about the nature of geometric objects that will also emerge in section 1.3. The result was Euclid's *Elements*, produced around 300 BCE, a system of proof that placed geometry on a firm (but wordy) logical foundation. Chapter 2 expands figure 1.2 into a proof in the style of Euclid. We will see that the saying "a picture is worth a thousand words" is pretty close to the mark.

Origins of the Pythagorean Theorem

As noted above, the Pythagorean theorem was discovered independently in several ancient cultures, probably earlier than Pythagoras himself. Special cases of it occur in ancient India and China, and perhaps earliest of all in Babylonia (part of modern Iraq). Thus the theorem is a fine example of the universality of mathematics. As we will see in later chapters, it recurs in different guises throughout the history of geometry, and also in number theory.

It is not known how it was first proved. The proof above is one suggestion, given by Heath (1925, 1:354) in his edition of the *Elements*. The Chinese and Indian mathematicians were more interested in triangles whose sides had particular numerical values, such as 3, 4, 5 or 5, 12, 13.

As we will see in the next section, the Babylonians developed the theory of numerical right-angled triangles to an extraordinarily high level.

1.2 PYTHAGOREAN TRIPLES

If the sides of a right-angled triangle are a, b, c, with c the hypotenuse, then the Pythagorean theorem is expressed by the equation

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

in the algebraic notation of today. Indeed, we call a^2 "a squared" in memory of the fact that a^2 represents a square of side a. We also understand that a^2 is found by multiplying a by itself, and the Pythagoreans would have agreed with us when a is a whole number. What made the Pythagorean theorem interesting to them are the whole-number triples $\langle a, b, c\rangle$ satisfying the equation above. Today, such triples are known as **Pythagorean triples**. The simplest example is of course $\langle 3, 4, 5\rangle$, because

$$3^2 + 4^2 = 9 + 16 = 25 = 5^2,$$

but there are infinitely many Pythagorean triples. In fact, the right-angled triangles whose sides are Pythagorean triples come in infinitely many shapes because the slopes b/a of their hypotenuses can take infinitely many values.

The most impressive evidence for this fact appears on a Babylonian clay tablet from around 1800 BCE. The tablet, known as Plimpton 322 (its catalog number in a collection at Columbia University), contains columns of numbers that Neugebauer and Sachs (1945) interpreted as values of b and c in a table of Pythagorean triples. Part of the tablet is broken off, so what remains are pairs $\langle b, c\rangle$ rather than triples. Some have questioned whether the Babylonian compiler of the tablet really had right-angled triangles in mind. In my opinion, yes, because all the values $c^2 - b^2$ are perfect squares *and* the pairs $\langle b, c\rangle$ are listed in order of the values b/a—the slopes of the corresponding hypotenuses. Figure 1.3 is a completed table that includes the values of a and b/a and also a fraction x that I explain below.

The column of a values reveals something else interesting. These values are all divisible only by powers of 2, 3, and 5, which makes them particularly "round" numbers in the Babylonian system, which was based on the number 60 (some of their system survives today, with 60 minutes in a hour and 60 seconds in a minute).

We do not know how the Babylonians discovered these triples. However, the amazingly complex values of b and c can be generated from the

a	b	c	b/a	x
120	119	169	0.9917	12/5
3456	3367	4825	0.9742	64/27
4800	4601	6649	0.9585	75/32
13500	12709	18541	0.9414	125/54
72	65	97	0.9028	9/4
360	319	481	0.8861	20/9
2700	2291	3541	0.8485	54/25
960	799	1249	0.8323	32/15
600	481	769	0.8017	25/12
6480	4961	8161	0.7656	81/40
60	45	75	0.7500	2
2400	1679	2929	0.6996	48/25
240	161	289	0.6708	15/8
2700	1771	3229	0.6559	50/27
90	56	106	0.6222	9/5

Figure 1.3 : Pythagorean triples in Plimpton 322

fractions x, which are fairly simple combinations of powers of 2, 3, and 5. In terms of x, the whole numbers a, b, and c are denominator and numerators of the fractions

$$\frac{b}{a} = \frac{1}{2}\left(x - \frac{1}{x}\right) \quad \text{and} \quad \frac{c}{a} = \frac{1}{2}\left(x + \frac{1}{x}\right).$$

For example, with $x = 12/5$ we get

$$\frac{1}{2}\left(x - \frac{1}{x}\right) = \frac{1}{2}\left(\frac{12}{5} - \frac{5}{12}\right) = \frac{119}{120} \quad \text{and} \quad \frac{1}{2}\left(x + \frac{1}{x}\right) = \frac{1}{2}\left(\frac{12}{5} + \frac{5}{12}\right) = \frac{169}{120}.$$

The huge triple $\langle 13500, 12709, 18541\rangle$ is similarly generated from the fraction $125/54 = 5^3/2 \cdot 3^3$, which has roughly the same complexity as $13500 = 2^2 \cdot 3^3 \cdot 5^3$. Thus, it is plausible that the Babylonians could have generated complex Pythagorean triples by relatively simple arithmetic. At the same time, the link with geometry is hard to deny when the triples are seen to be arranged in order of the slopes b/a—an order that could not be guessed from the arrangement of a, b, c, or x values! And when one sees that these slopes cover a range of angles, roughly equally spaced, between 30° and 45° (figure 1.4), it looks as though the Babylonians were collecting triangles of different shapes.

It is also conspicuous which shape is *missing* from this collection of triangles: the one with equal sides a and b, shown in red in figure 1.4.

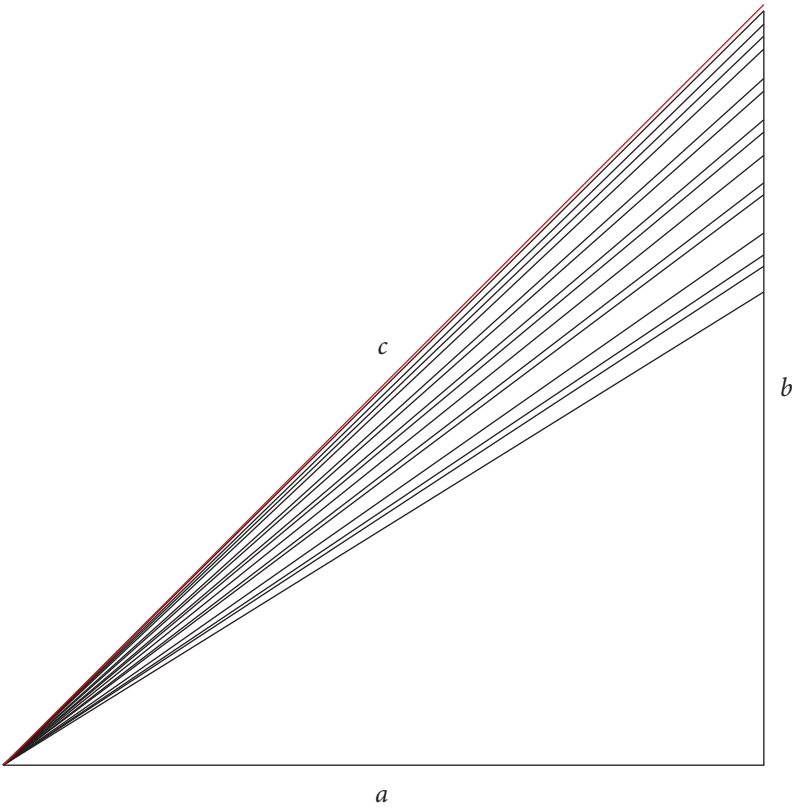


Figure 1.4 : Slopes derived from Plimpton 322

As we now know, because the Pythagoreans discovered it, this shape is missing because the hypotenuse of this triangle is *irrational*.

1.3 IRRATIONALITY

Irrationality follows naturally from the Pythagorean theorem, but apparently it was found by the Pythagoreans alone. Like other discoverers of the theorem, the Pythagoreans knew special cases with whole-number values of a, b, c. But, apparently they were the only ones to ask, Why do we find no such triples with $a = b$? The question points to its own answer: *it is contradictory to suppose there are whole numbers a and c such that* $c^2 = 2a^2$.

The argument of the Pythagoreans is not known, but the result must have been common knowledge by the time of Aristotle (384–322 BCE),

as he apparently assumes his readers will understand the following brief hint:

> The diagonal of the square is incommensurable with the side, because odd numbers are equal to evens if it is supposed commensurable.
>
> (Aristotle, *Prior Analytics*, bk. 1, chap. 23)

Here "commensurable" means being a whole number multiple of a common unit of measure, so we are supposing that $c^2 = 2a^2$, where the side of the square is a units and its diagonal is c units. We reach the contradiction "odd = even" as follows.

First, by choosing the unit of measure as large as possible, we can assume that the whole numbers c and a have no common divisor (except 1). In particular, at most one of them can be even.

Now $c^2 = 2a^2$ implies that the number c^2 is even. Since the square of an odd number is odd, c must also be even, say $c = 2d$. Substituting $2d$ for c gives

$$(2d)^2 = 2a^2 \quad \text{so} \quad 2d^2 = a^2.$$

But then a similar argument shows a is even, which is a contradiction.

So it is wrong to suppose there are whole numbers a and c with $c^2 = 2a^2$.

The usual way to express this fact today is that *there are no natural numbers c and a such that* $\sqrt{2} = c/a$ or, more simply, that $\sqrt{2}$ *is irrational.*

1.4 FROM IRRATIONALS TO INFINITY

The argument for irrationality of $\sqrt{2}$ is very short and transparent in modern algebraic symbolism. Judging by the excerpt from Aristotle, it was also comprehensible enough when equations were written out in words, as the ancient Greeks did.

But there was also a geometric approach to incommensurable quantities that the Greeks called *anthyphaeresis*. It gives a different and deeper insight into the nature of $\sqrt{2}$ and, indeed, a different proof that it is irrational. Anthyphaeresis is a process that can be applied to two quantities, such as lengths or natural numbers, by repeatedly subtracting the smaller from the larger. Since it was later used to great effect by Euclid, it is today called the **Euclidean algorithm**.

More formally, given two quantities a_1 and b_1 with $a_1 > b_1$, one forms the new pair of quantities b_1 and $a_1 - b_1$ and calls the greater of them a_2

and the lesser b_2. Then one does the same with the pair a_2, b_2, and so on. For example, if $a_1 = 5$, $b_1 = 3$ we get

$$\langle a_1, b_1 \rangle = \langle 5, 3 \rangle$$
$$\langle a_2, b_2 \rangle = \langle 3, 2 \rangle$$
$$\langle a_3, b_3 \rangle = \langle 2, 1 \rangle$$
$$\langle a_4, b_4 \rangle = \langle 1, 1 \rangle,$$

at which point the algorithm terminates because $a_4 = b_4$. The Euclidean algorithm always terminates when a_1 and b_1 are natural numbers, because subtraction produces smaller natural numbers and natural numbers cannot decrease forever. Conversely, *a ratio for which the Euclidean algorithm runs forever is irrational*.

In section 2.6 we will see the consequences of the Euclidean algorithm for natural numbers, but for the Greeks before Euclid the process of anthyphaeresis was most revealing for pairs of incommensurable quantities, such as $a_1 = \sqrt{2}$ and $b_1 = 1$. In this case the numbers a_n, b_n can and do decrease forever. In fact, we have

$$\langle a_1, b_1 \rangle = \langle \sqrt{2}, 1 \rangle$$
$$\langle a_2, b_2 \rangle = \langle 1, \sqrt{2} - 1 \rangle$$
$$\langle a_3, b_3 \rangle = \langle 2 - \sqrt{2}, \sqrt{2} - 1 \rangle = \langle (\sqrt{2} - 1)\sqrt{2}, (\sqrt{2} - 1)1 \rangle,$$

so $\langle a_3, b_3 \rangle$ is the same as $\langle a_1, b_1 \rangle$, just scaled down by the factor $\sqrt{2} - 1$. Two more steps will give $\langle a_5, b_5 \rangle$, again the same as $\langle a_1, b_1 \rangle$ but scaled down by the factor $(\sqrt{2} - 1)^2$, and so on. Thus the numbers $\langle a_n, b_n \rangle$ decrease forever, but they return to the same ratio every other step.

Since this cannot happen for any pair $\langle a, b \rangle$ of natural numbers, it follows that $\sqrt{2}$ and 1 are not in a natural number ratio; that is, $\sqrt{2}$ is irrational. Moreover, we have discovered that the pair $\langle \sqrt{2}, 1 \rangle$ behaves *periodically* under anthyphaeresis, producing pairs in the same ratio every other step. It turns out, though this was not understood until algebra was better developed, that periodicity is a special phenomenon occurring with square roots of natural numbers.

Visual Form of the Euclidean Algorithm

If a and b are lengths, we can represent the pair $\{a, b\}$ by the rectangle with adjacent sides a and b. If, say, $a > b$, then the pair $\{b, a - b\}$ is represented by the rectangle obtained by cutting a square of side b from the

original rectangle, shown in light gray in figure 1.5. The algorithm then repeats the process of cutting off a square in the light gray rectangle, and so on.

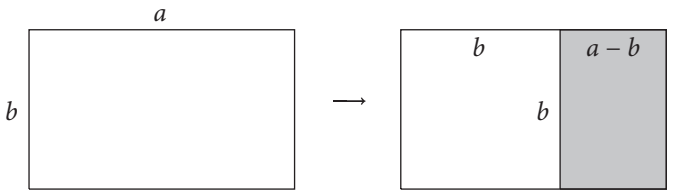


Figure 1.5 : First step of the Euclidean algorithm

When $a = \sqrt{2}$ and $b = 1$, two steps of the algorithm give the light gray rectangle shown in figure 1.6, which is the *same shape* as the original rectangle. This is because its sides are again in the ratio $\sqrt{2} : 1$, as we saw in the calculation above. Since the new rectangle is the same shape as the old, it is clear that the process of cutting off a square will continue forever.

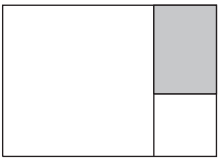


Figure 1.6 : After two steps of the algorithm on $\sqrt{2}$ and 1

The Greeks were fascinated by geometric constructions in which the original figure reappears at a reduced size. The simplest example is the so-called *golden rectangle* (see figure 1.7), in which removal of a square leaves a rectangle the same shape as the original. It follows that the Euclidean algorithm runs forever on the sides a and b of the golden rectangle, and hence these sides are in irrational ratio. This particular ratio is called the **golden ratio**.

Figure 1.7 : The golden rectangle

The golden ratio is also the ratio of the diagonal to the side of the regular pentagon, where the recurrence of the original figure at reduced size can be seen in figure 1.8.

It is believed that the study of the golden ratio and the regular pentagon may go back to the Pythagoreans, in which case they were probably aware of the irrationality of the golden ratio as well as that of $\sqrt{2}$.

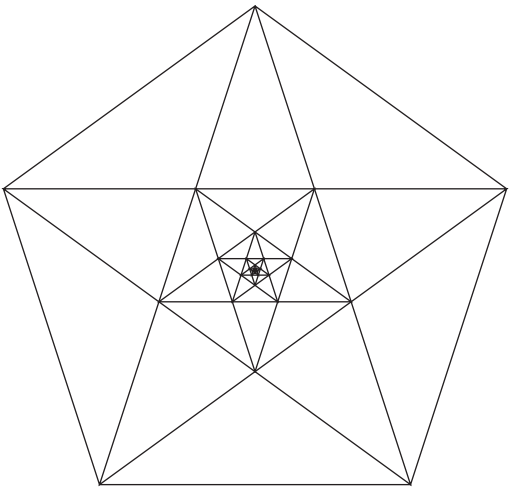


Figure 1.8 : Infinite series of pentagrams

1.5 FEAR OF INFINITY

As we have just seen, irrationality brings infinite processes to the attention of mathematicians, albeit processes of a simple and repetitive kind. At an even more primitive level, the natural numbers $0, 1, 2, 3, \ldots$ themselves represent the kind of infinity where a simple process—in this case, adding 1—is repeated without end. An infinity that involves endless repetition was called by the Greeks a *potential infinity*. They contrasted it with *actual infinity*—a somehow completed infinite totality—which was considered unacceptable or downright contradictory.

The legendary opponent of infinity was Zeno of Elea, who lived around 450 BCE. Zeno posed certain "paradoxes of the infinite," which we know only from Aristotle, who described the paradoxes only to debunk them, so we do not really know what Zeno meant by them or how they

were originally stated. It will become clear, however, that Zeno accepted potential infinity while rejecting actual infinity.

A typical Zeno paradox is his first, the *paradox of the dichotomy*, in which he argues that motion is impossible because

> before any distance can be traversed, half the distance must be traversed [and so on], that these half distances are infinite in number, and that it is impossible to traverse distances infinite in number. (Aristotle, *Physics*, bk. 8, chap. 8, 263a)

Apparently, Zeno is arguing that the infinite sequence of events

reaching 1/2 way
reaching 1/4 way
reaching 1/8 way
…

cannot be completed. Aristotle answers, a few lines below this statement, that

> the element of infinity is present in the time no less than in the distance.

In other words, if one can conceive an infinite sequence of places

$$1/2 \text{ way}, 1/4 \text{ way}, 1/8 \text{ way}, \ldots.$$

then one can conceive an infinite sequence of times at which

$$1/2 \text{ way is reached}, 1/4 \text{ way is reached}, 1/8 \text{ way is reached}, \ldots.$$

Thus if Zeno is willing to admit the potential infinity of places, he has to admit the potential infinity of times. It is not a question of *completing* an infinity but only of correlating one potential infinity with another. We claim only that each of the places can be reached at a certain time; we do not have to consider the totality of places or the totality of times.

At any rate, after Zeno, Greek mathematicians handled questions about infinity by this style of argument—dealing with members of a potential infinity one by one rather than in their totality. The "actual infinity scare" was nevertheless productive, because it led to a very subtle understanding of the relation between the continuous and the discrete.

1.6 EUDOXUS

Eudoxus of Cnidus, who lived from approximately 390 BCE to 330 BCE, was a student of Plato and is believed to have taught Aristotle. His most important accomplishments are the **theory of proportions** and the **method of exhaustion**. Together, they form the summit of the Greek treatment of infinity, and they come down to us mainly through the exposition in book 5 of Euclid's *Elements*. In particular, the theory of proportions was the best treatment of rational and irrational quantities available until the nineteenth century. Indeed, it is probably the best treatment possible as long as one rejects actual infinity, which most mathematicians did until the 1870s.

The theory of proportions deals with "magnitudes" (typically lengths) and their relation to "numbers," which are natural numbers. It thereby builds a bridge between the two worlds separated by the Pythagoreans: the world of magnitudes, which vary *continuously*, and the world of counting, where numbers jump *discretely* from each number to its successor.

The theory is complicated somewhat because the Greeks thought in terms of ratios of magnitudes and ratios of numbers, without having the algebraic machinery of fractions that makes ratios easy to handle. We can understand the ratio of natural numbers m and n as the fraction m/n, so we will write the ratio of lengths a and b as the fraction a/b.[1] The key idea of Eudoxus is that ratios of lengths, a/b and c/d, are equal if and only if, for *each* natural number ratio m/n,

$$\frac{m}{n} < \frac{a}{b} \quad \text{if and only if} \quad \frac{m}{n} < \frac{c}{d}.$$

Equivalently (and this is how Eudoxus put it), for each natural number pair m and n,

$$mb < na \quad \text{if and only if} \quad md < nc.$$

Thus the infinity of natural number pairs m, n is behind the definition of equality of length ratios, but only potentially so, because equality

1. It may seem unwieldy to work with ratios of lengths rather than just lengths, but in fact length is a *relative* concept and only the ratio of lengths is absolute. When we say length $a = 3$, for example, we really mean that 3 is the ratio of a to the unit length. In chapter 9 we will see that the relative concept of length is a specific characteristic of Euclidean geometry.

depends on a single (though arbitrary) pair m, n. In defining unequal length ratios, infinity can be avoided completely, because one *particular* pair can witness inequality. Namely, if $a/b < c/d$ then there is a particular m/n such that

$$\frac{a}{b} < \frac{m}{n} < \frac{c}{d},$$

and likewise, if $c/d < a/b$ then there is a particular m/n between c/d and a/b. Today we would say that ratios of lengths are *separable* by ratios of natural numbers.

The Archimedean Axiom

The assumption that natural number ratios separate ratios of lengths is equivalent to a property later called the *Archimedean property*: if $a/b > 0$ then $a/b > m/n > 0$ for some natural numbers m and n. It follows, obviously, that in fact $a/b > 1/n$, so $na > b$. This gives the usual statement of the **Archimedean axiom**: *if a and b are any nonzero lengths, then there is a natural number n such that* $na > b$.

Another statement of the Archimedean axiom is: *there is no ratio* a/b *so small that* $0 < a/b < 1/n$ *for each natural number n*, or more concisely, *there are no infinitesimals*. This property was assumed by Euclid and Archimedes (hence the name), but some later mathematicians, such as Leibniz, thought that infinitesimals exist. We will see in chapter 4 that the existence of infinitesimals was a big issue in the development of calculus.

Mathematical practice today has translated Eudoxus's theory into our concept of the **real number system** $\mathbb{R}$. The ratios of lengths are the nonnegative real numbers, and among them lie the nonnegative **rational numbers**, which are the ratios m/n of natural numbers. Any two distinct real numbers are separated by a rational number, so there are no infinitesimals in $\mathbb{R}$. Conversely, each real number is determined by the rational numbers less than it and the rational numbers greater than it. Exactly how this came about, and what the real numbers *are*, is explained in chapter 11. It turns out that separation by rational numbers is the key to answering this question.

The Method of Exhaustion

We discuss the method of exhaustion only briefly here, because it is a generalization of the theory of proportions. Also, the best examples of the method occur in the work of Euclid and Archimedes, discussed in

chapter 2. The basic idea is to approximate an "unknown quantity," such as the area or volume of a curved region, by "known quantities" such as areas of triangles or volumes of prisms. This generalizes the idea of approximating a ratio of lengths by ratios of natural numbers. Generally, there is a potential infinity of approximating objects, but as long as they come "arbitrarily close" to the unknown quantity it is possible to draw conclusions without appealing to actual infinity.

An example is approximation of the circle by polygons, shown in figure 1.9, which allows us to draw the conclusion that the area of the circle is proportional to the square of its radius.

Figure 1.9 shows polygons approximating the circle from inside and outside. Only the first two approximations are shown, but one can imagine a continuation of the sequence by repeatedly doubling the number of sides. It is clear that the area of the gap between inner and outer polygons becomes arbitrarily small in the process, and hence both inner and outer polygons come arbitrarily close to the circle in area.

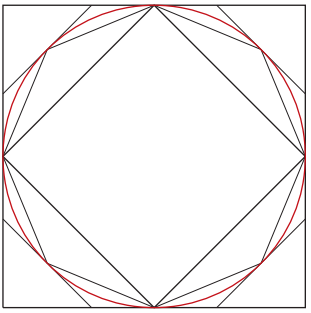


Figure 1.9 : Approximating the circle by polygons

Also, the area of each polygon P_n is a sum of triangles, whose area $P_n(R)$ for radius R is known and proportional to R^2. Now comes a typical example of reasoning "by exhaustion": suppose that the area $C(R)$ of the circle of radius R is *not* proportional to R^2. Thus, if we compare circles of radius R and R' we have either

$$C(R)/C(R') < R^2/R'^2$$

or

$$C(R)/C(R') > R^2/R'^2.$$

If $C(R)/C(R') < R^2/R'^2$, then by choosing n so that $P_n(R)$ is sufficiently close to $C(R)$ and $P_n(R')$ is sufficiently close to $C(R')$, we will get

$$P_n(R)/P_n(R') < R^2/R'^2,$$

which is a contradiction. If $C(R)/C(R') < R^2/R'^2$ we get a similar contradiction. Therefore *the only possibility is that* $C(R)/C(R') = R^2/R'^2$.

We have established what we want by *exhausting* all other possibilities. This is what "exhaustion" means in the method of exhaustion. Notice also that we used only the potential infinity of polygons by going only far enough to contradict a given inequality. This is typical of the method.

1.7 REMARKS

We have seen in the development of Greek mathematics many topics considered tricky in undergraduate mathematics today, such as proof by contradiction, the use of infinity, and the idea of choosing a "sufficiently close" approximation. This just goes to show, in my opinion, that ancient mathematics is good training in the art of proof.

At the same time, we have seen that ancient arguments can often be streamlined by the use of algebraic symbolism, and the art of algebra was missing in ancient times.

The other thing missing, in what we know of this early stage, was the systematic deduction of theorems from axioms. The art of **axiomatics** also began in ancient times, as we will see in the next chapter.

CHAPTER 2

Euclid

In the story of proof, Euclid comes near the beginning because his *Elements* was composed around 300 BCE and few earlier examples of proof survive. Unfortunately, this means plunging the reader into deep water immediately, because Euclid did so much that the *Elements* became the model of proof until quite recent times. There was no major advance in the technique of proof until algebraic symbolism was added in the sixteenth century, and no advance in logic itself until the nineteenth century.

Also, the *Elements* is conceptually subtle in separating the continuous (geometry) from the discrete (number theory), following the Pythagorean separation of quantity and "number." The difficult book 5 begins to build a bridge between the two, with the theory of proportions, and by admitting (limited) use of infinity. Infinity is also used in an elegant determination of the volume of the tetrahedron.

Because of the many difficulties in the *Elements*, readers may prefer to skim the next two chapters and come back to the details later. Still, *some* acquaintance with Euclid is needed to understand the later development of mathematics. The *Elements* influenced not only mathematics but also philosophy (Spinoza's *Ethics*) and law (Abraham Lincoln was an admirer). However, philosophy and law could not attain a standard of proof *higher* than that of the *Elements*, whereas mathematics could.

Eventually, in the nineteenth century, mathematicians became aware of gaps in Euclid's reasoning and of alternatives to his axioms, which led to more rigorous foundations of mathematics by the end of the

nineteenth century. At this point another "crisis of foundations" emerged and transformed the concept of proof in many ways, some of which are still being worked out.

2.1 DEFINITION, THEOREM, AND PROOF

Euclid's *Elements* is the oldest mathematics book that looks "modern," in the sense of containing definitions, theorems, and proofs, arranged in logical order. On closer inspection one sees some flaws—Euclid tries to define terms that should remain undefined, and he tries to prove some statements that should be axioms—but nevertheless, the *Elements* is a masterpiece that set the standard of mathematical proof for over 2,000 years. Perhaps the most important lesson taught by the *Elements* is that mathematics can be built, cumulatively, by deduction from self-evident *axioms*.

The *Elements* is founded on simple objects such as points, lines, and circles, the associated quantities of length and angle, and certain axioms about them (which are traditionally called *postulates*). These axioms are, in the classic translation of Heath (1925):

P1. To draw a straight line from any point to any point.
P2. To produce a finite straight line continuously in a straight line.
P3. To describe a circle with any center and distance.
P4. That all right angles are equal to one another.
P5. That, if a straight line falling on two straight lines make 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.

One sees immediately a peculiarity of Euclid's language, which favors **construction** over mere **existence**. Postulate 1 does not say there *exists* a straight line (segment) between any two points but, rather, that the line segment *can be drawn*. And Postulate 2 does not say a straight line is infinite but (more conservatively) that a line segment can be *produced continuously*, that is, extended indefinitely. The question of construction is an important secondary theme of the *Elements*, and many of his theorems state that a certain figure can be constructed by the instruments that draw straight lines and circles ("straightedge" and "compass"). Unfortunately, Euclid's very first construction, on which several others

depend, does *not* follow from his axioms. We will therefore assume existence in cases where Euclid makes a construction and postpone the question of extra axioms until later.

Perhaps the oddest postulate, to modern ears, is "all right angles are equal." To make sense of this, one has to realize that for Euclid an angle is merely a pair of half lines with a common endpoint—it does not come with a measure in degrees or radians. One can only say whether angles are equal or not, and a right angle ABC is one for which the two angles ABC and ABD in figure 2.1 are equal. Postulating that all right angles are equal then gives a standard unit of angle measure, *the* right angle. Indeed, one finds throughout the *Elements* that angles (or sums of angles) are always given as multiples of the right angle.

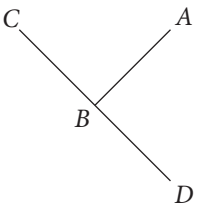


Figure 2.1 : Right angles

Postulate 5, known as the **parallel axiom**, actually states the condition for lines *not* to be parallel. If line $\mathscr{N}$ falls on lines $\mathscr{L}$ and $\mathscr{M}$ and makes angles α and β as shown in figure 2.2, and if $\alpha + \beta$ is less than two right angles, then postulate 5 says that $\mathscr{L}$ and $\mathscr{M}$ will meet somewhere on the right.

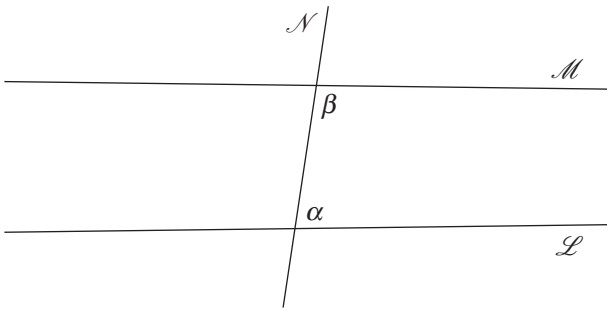


Figure 2.2 : Nonparallel lines

Therefore, if $\mathscr{L}$ and $\mathscr{M}$ do *not* meet—that is, are parallel—then $\alpha + \beta$ equals two right angles. This is one of many equivalent statements of the parallel axiom, convenient for proving that the angle sum of any triangle is two right angles. There are also equivalents of the axiom that do not mention the concept of angle, for example, the so-called **Playfair's axiom** saying that for any line $\mathscr{L}$ and a point P outside it there is exactly one line $\mathscr{M}$ through P that does meet $\mathscr{L}$ (see section 2.3).

As we will see later, many mathematicians were dissatisfied with the parallel axiom and hoped that it could be proved from postulates 1–4. However, this turned out to be impossible, for very interesting reasons we will see in chapter 9.

The principles of deduction were not explicitly stated in the *Elements*, except for the following, which Euclid called *common notions*. They can be viewed as properties of equality (and inequality), addition, and subtraction. (The first four were used in section 1.1 for a visual proof of the Pythagorean theorem.)

Common notion 1. Things that are equal to the same thing are equal to each other.
Common notion 2. If equals are added to equals, the wholes are equal.
Common notion 3. If equals are subtracted from equals, the remainders are equal.
Common notion 4. Things that coincide with one another are equal.
Common notion 5. The whole is greater than the part.

When the common notions are written in modern symbolism, they look rather like principles of algebra:

1. If $A = B$ and $C = B$ then $A = C$.
2. If $A = B$ then $A + C = B + C$.
3. If $A = B$ then $A - C = B - C$.
4. $A = A$.
5. If $A \subset B$ then $A < B$.

However, despite this promising start, algebra failed to materialize in the *Elements*. We take up the question of algebra again later. Now let us look at Euclid's theorems, or "propositions" as they are traditionally called. It is enough to look at book 1, which already contains some remarkable deductions.

2.2 THE ISOSCELES TRIANGLE THEOREM AND SAS

As a simple example of a deduction in Euclid's system, let us show that his proposition 5 follows from his proposition 4. Concisely stated, these propositions in book 1 of the *Elements* are:

Proposition 4. If two triangles agree in two sides and the included angle, then they agree in all corresponding sides and angles.

Proposition 5. If a triangle has two equal sides, then its two angles, other than the angle included by the equal sides, are also equal.

In modern geometry, proposition 4 is often abbreviated SAS, for "side angle side," and is considered an axiom. The triangle described in proposition 5 is called *isosceles*, from the Greek for "equal sides." Euclid's proof that SAS implies the isosceles triangle theorem was a traditional stumbling block for students of the *Elements*, known as the "ass's bridge" because asses could not get past it (or possibly because of Euclid's diagram, which consists of five lines resembling a bridge).

A much shorter proof was given by the later Greek geometer Pappus, so let's look at the Pappus proof instead. The reader should be warned, however, that the Pappus proof is almost too clever, because it takes the two triangles in SAS to be the *same* triangle. This is OK because no one said that the two triangles have to be different!

Suppose, then, that ABC is a triangle with $AB = AC$, as in figure 2.3. Notice that we may also take this to be a picture of triangle ACB.

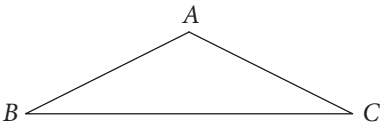


Figure 2.3 : An isosceles triangle

Now, the triangles ABC and ACB agree in two sides and the included angle, because $AB = AC$, $AC = AB$, and the angle at A is the included angle in both. Therefore, by proposition 4, the triangles agree in all corresponding angles. In particular, the angle at B (in triangle ABC) equals the angle at C (in triangle ACB, which of course is the same triangle). That's it!

SAS Implies ASA

SAS states a condition for triangles to be *congruent*—meaning they agree in all side lengths and all corresponding angles. Another such condition is ASA (for "angle side angle"): *if triangles agree in two angles and the side common to these angles, then they are congruent*. ASA is part of Euclid's proposition 26. It follows from SAS by a logical device already used in section 1.3: prove a statement false by showing that it leads to contradiction.

Suppose that $A_1B_1C_1$ and $A_2B_2C_2$ are two triangles, with equal angles α and β as shown in figure 2.4, and $A_1B_1 = A_2B_2$. Thus $A_1B_1C_1$ and $A_2B_2C_2$ agree in two angles and their common side, so ASA holds. Now suppose, for the sake of contradiction, that $A_1B_1C_1$ and $A_2B_2C_2$ are *not* congruent.

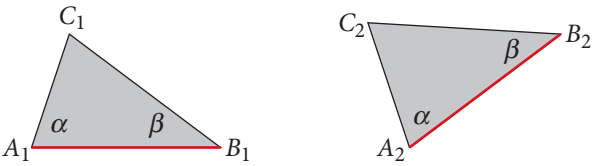


Figure 2.4 : Triangles satisfying ASA

Then not all corresponding sides are equal, else SAS holds and the triangles are congruent, contrary to our supposition. So, some corresponding sides are *un*equal, and (renaming if necessary) we can assume $A_1C_1 < A_2C_2$.

But then we can choose a point C on A_2C_2, between A_2 and C_2, so that $A_2C = A_1C_1$. Hence drawing the line B_2C creates an angle β' that is only part of β (figure 2.5), so $\beta > \beta'$ because "the whole is greater than the part."

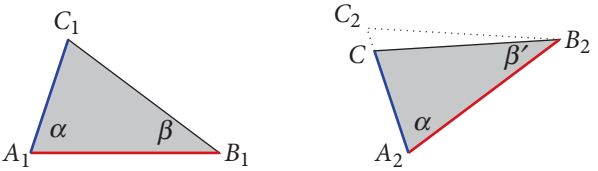


Figure 2.5 : Hypothetical triangles satisfying SAS

Yet the triangles $A_1B_1C_1$ and A_2B_2C satisfy SAS, since $A_2C = A_1C_1$, and hence they are congruent. In particular, β' in A_2B_2C equals the corresponding angle β in $A_1B_1C_1$, which is again a contradiction.

Therefore, it is false to suppose $A_1B_1C_1$ and $A_2B_2C_2$ are not congruent.

2.3 VARIANTS OF THE PARALLEL AXIOM

The proof of ASA above has the bonus feature that it holds even if the point C_2 does not exist! That is, we need only assume that the second "triangle" consists of the segment A_2B_2 and lines out of A_2 and B_2 at angles α and β, respectively. On the line through A_2 we can still choose the point C so that $A_2C = A_1C_1$ and arrive a contradiction as above.

This strong version of ASA enables us to prove the following variant of the parallel axiom P5: *If a straight line $\mathscr{N}$ falling on two straight lines $\mathscr{L}$ and $\mathscr{M}$ makes angles α and β, respectively, on the same side, with $\alpha + \beta =$ two right angles, then $\mathscr{L}$ and $\mathscr{M}$ are parallel.*

Suppose we have a line $\mathscr{N}$ that crosses two lines $\mathscr{L}$ and $\mathscr{M}$, making angles α and β as shown in figure 2.6, so $\alpha + \beta =$ two right angles.

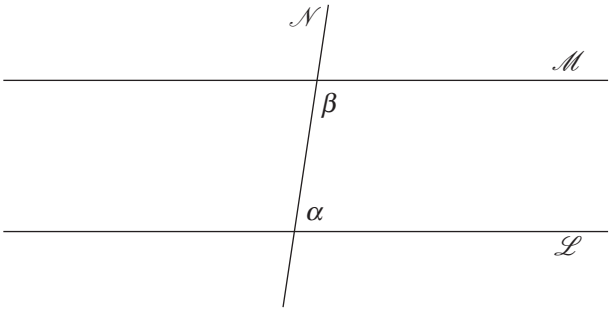


Figure 2.6 : Parallel lines

We can then find all the angles in figure 2.6 with the help of Euclid's book 1, proposition 13, which states: *If α and β together make a straight angle, then $\alpha + \beta =$ two right angles*. This proposition is proved by considering two angles α and β that make a straight angle at point P and comparing them with two right angles ρ that meet at P (figure 2.7). Since all right angles are equal, we can call them all ρ and, by subtraction, get the three angles shown on the right of figure 2.7. Since the right angle on

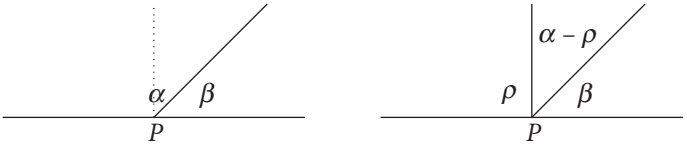


Figure 2.7 : Lines making a straight angle

the right consists of $\alpha - \rho$ and β, we get $\rho = \alpha - \rho + \beta$ and therefore

$$\alpha + \beta = 2\rho.$$

It follows from proposition 13, by "subtracting equals from equals," that the angles α and β formed by $\mathscr{L}$, $\mathscr{M}$, and $\mathscr{N}$ in figure 2.6 recur as shown in figure 2.8.

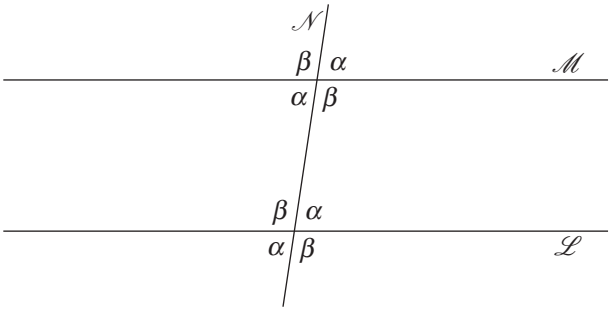


Figure 2.8 : Angles related to parallels

However, we do not yet know that $\mathscr{L}$ and $\mathscr{M}$ are parallel! We prove that they are with the help of the strong ASA. If in fact the lines $\mathscr{L}$ and $\mathscr{M}$ in figure 2.8 meet (say, on the right), then they form a triangle in combination with the segment of $\mathscr{N}$ between them. The same segment and angles occur on the left, forming a triangle congruent to the one on the right, by ASA. But then the lines $\mathscr{L}$ and $\mathscr{M}$ meet at two points—one on the right and one on the left. This is contrary to axiom P1, which, although Euclid does not say so explicitly, gives a *unique* line between any two points.

This contradiction shows that $\mathscr{L}$ and $\mathscr{M}$ do not meet; that is, they are parallel.

It follows that *there is a unique parallel $\mathscr{M}$ to a given line $\mathscr{L}$ through a given point P outside $\mathscr{L}$*. This is because for any such *P* we can choose a

line $\mathscr{N}$ through P that crosses $\mathscr{L}$ at angle α, say. We then choose the line $\mathscr{M}$ through P that crosses $\mathscr{N}$ at angle β, where $\alpha + \beta$ = two right angles.

The italicized sentence above is another equivalent of the parallel axiom, often the most convenient one, because it avoids the concept of angle and it says that parallels exist. It is called **Playfair's axiom**, after the Scottish mathematician John Playfair (1748–1819). It appeared in his book Playfair (1795).

Parallelograms and Triangles

With the existence of parallels we get the existence of *parallelograms*, the four-sided figures whose opposite sides are parallel. Figure 2.8 gives us some equal angles. Figure 2.9 shows some of the equal angles, all of which can be deduced from figure 2.8 by choosing which lines to interpret as $\mathscr{L}$, $\mathscr{M}$, and $\mathscr{N}$.

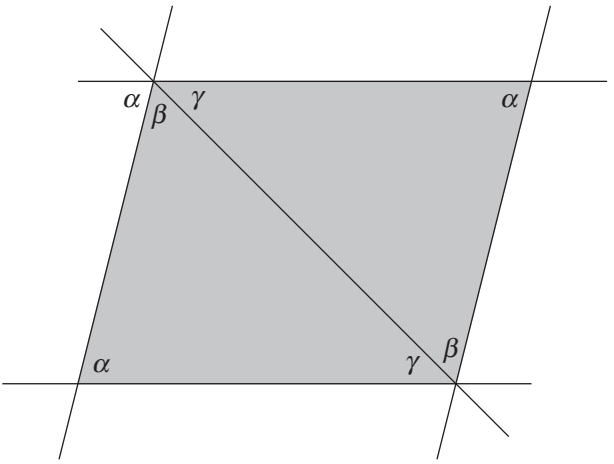


Figure 2.9 : Angles associated with a parallelogram

Notice that the gray parallelogram consists of two triangles with the diagonal as their common side and corresponding equal angles β and γ. By ASA, these triangles are congruent; hence *opposite sides of a parallelogram are equal*. Notice also in figure 2.9 that the angle sum $\alpha + \beta + \gamma$ of each triangle equals the straight angle at the top left, and therefore *the angle sum of any triangle is two right angles*. The latter statement is actually Euclid's proposition 32 of book 1. It too is an equivalent of the parallel axiom.

To sum up, we have found the following consequences of Euclid's parallel axiom (which in fact also imply it, and hence are equivalent to it):

Playfair's axiom. For any line $\mathscr{L}$ and a point P outside it, there is a unique parallel to $\mathscr{L}$ through P.

Angle sum of a triangle. The angles of any triangle have sum equal to two right angles.

2.4 THE PYTHAGOREAN THEOREM

> He was 40 years old before he looked on Geometry; which happened accidentally. Being in a Gentleman's Library, Euclid's Elements lay open, and 'twas the 47 El. libri I. He read the Proposition. *By* G—sayd he (he would now and then sweare an emphaticall Oath by way of emphasis) *this is impossible*! So he reads the Demonstration of it, which referred him back to such a Proposition; which proposition he read. That referred him back to another, which he also read …that at last he was demonstratively convinced of that trueth. This made him in love with Geometry.

This quotation is about the philosopher Thomas Hobbes (1588–1679), from *Brief Lives* by Aubrey (1898: 332). It neatly and concisely captures the effectiveness of the deductive method: by seeing how a proposition depends on previous propositions, and ultimately on self-evident propositions called *axioms*, anyone can be convinced of its truth. As long as each proposition is a logical consequence of the previous propositions, and hence of the axioms, it does not matter how long and complex the chain of consequences may be. The proposition that so impressed Hobbes is the Pythagorean theorem: proposition 47 in book 1 of the *Elements*.

Aubrey's account, incidentally, describes how a proof should first be read, which is *backward*: first find what propositions the theorem depends on, and then observe how it follows from these propositions. In the process, one learns what concepts are involved and how they are connected. This is not to say that it is easy to construct a proof in the first place. In fact, the proof that Hobbes loved is incredibly complicated when analyzed in detail. It involves dozens of connections. However, if one knows enough connections, one can string them together to

make proofs. In the last two sections we have already seen most of the connections needed to prove the Pythagorean theorem.

Figure 2.10 shows again a figure from section 1.1, this time with some sides and angles labeled to guide the steps of a proof. The square on the left, with each side $a + b$, has inside it four copies of the right-angled triangle with perpendicular sides a, b and angles α and β opposite to them. The light gray region therefore has each side equal to the hypotenuse c of the triangle. Also, the angle γ at each corner of the light gray region makes a straight angle with the angles α and β. Therefore,

$$\alpha + \beta + \gamma = \text{ two right angles.}$$

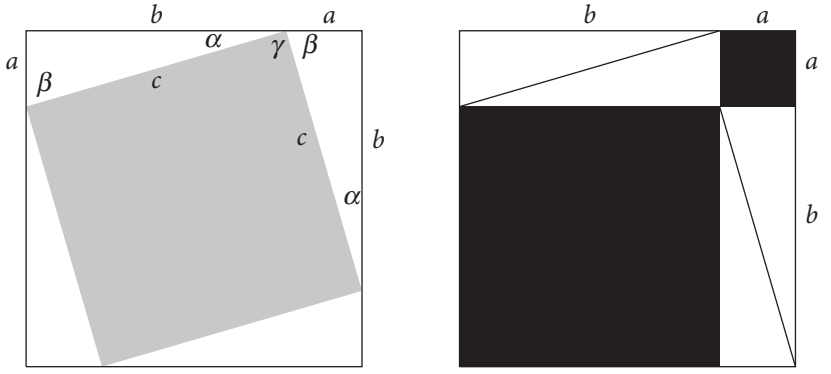


Figure 2.10 : Seeing how to prove the Pythagorean theorem

On the other hand, the angle sum of each triangle is also two right angles, so γ must be a right angle, and therefore the light gray region is the square on the hypotenuse. Thus the square on the hypotenuse equals the big square minus four times the triangle.

Turning now to the square on the right, which also has side $a + b$, we find similarly that the black regions are squares on the sides a and b. The sum of the black squares is again equal to the big square minus four times the triangle; hence it equals the square on the hypotenuse.

2.5 GLIMPSES OF ALGEBRA

As mentioned in section 2.1, Euclid's "common notions" look like algebra in the way they deal with equality, addition, and subtraction. Indeed, the proof above made extensive use of "adding equals to equals," "subtracting

equals from equals," and the principle that "things equal to the same thing are equal to each other." However, this is still not algebra as we know it because a full notion of **multiplication** is missing. Admittedly, we did say "four times the triangle," but this really meant

$$\text{triangle} + \text{triangle} + \text{triangle} + \text{triangle}.$$

We did not multiply one length (or area) by another.

This is because Euclid followed the Pythagoreans in denying geometric quantities such as length and area some of the attributes of numbers. Lengths can be added and subtracted, and we can decide whether two lengths are equal or not. But defining what **area** is and what it has to do with the product of lengths is already a complicated problem. Yet it is a problem we need to solve in order to understand what the "sum of two squares" means in the Pythagorean theorem. In the proof of the Pythagorean theorem we were able to solve it by showing the sum of two squares was equal to a single square by addition and subtraction of clearly equal areas.

Euclid solved the problem of defining area in general with great ingenuity, but unfortunately in a way that stymied the development of algebra in geometry until the seventeenth century.

Simply put, Euclid's solution was to define the product of line segments a and b to be the *rectangle* with adjacent sides a and b. This definition is compatible with multiplication when a and b are whole numbers—because the rectangle then consists of ab unit squares—but it is also meaningful when a or b is an irrational length, which the Greeks did not consider to be a number. The immediate difficulty with this definition is to decide **equality**; for example, is a rectangle with sides $\sqrt{2}$ and $\sqrt{3}$ equal to a rectangle with sides $\sqrt{6}$ and 1?

Before answering this question, let us consider some simple examples of polygons that might be considered "equal," by "addition" and "subtraction," according to Euclid's common notions 1, 2, and 3. First, a rectangle of width a and height b is equal to any parallelogram with the same base and height, as figure 2.11 suggests. This is because the rectangle results from the parallelogram by subtraction, then addition, of equal triangles.

The triangles are equal by the result from section 2.4 that opposite sides and opposite angles of a parallelogram are equal. Equality of parallelogram sides implies, by subtraction and addition again, that the

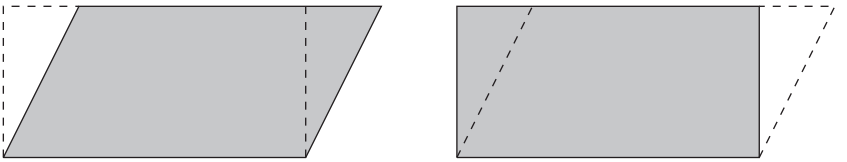


Figure 2.11 : Equal rectangle and parallelogram

triangles have equal width, and then equality of angles implies they are congruent, by SAS.

Next, one notices that any triangle, added to a copy of itself, makes a parallelogram (figure 2.12).

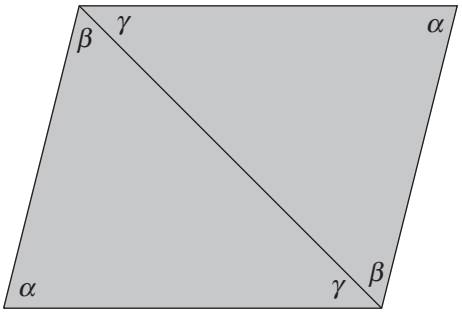


Figure 2.12 : Triangle and parallelogram

It follows that the area of a triangle is half that of a parallelogram with the same base and height, a result Euclid uses in his proof of the Pythagorean theorem. (This makes for a somewhat longer path to the theorem than the one described in the previous section.)

In general, Euclid considers regions "equal" if one can be converted to the other by addition and subtraction of finitely many equal figures. Remarkably, this definition coincides with the modern concept of "equal area" for polygons. However, the "product" *ab* has very limited algebraic properties. One has the *commutative law*

$$ab = ba,$$

because the rectangle with adjacent sides a and b is the *same* as the rectangle with adjacent sides b and a. And one has the *distributive law*

$$a(b + c) = ab + ac,$$

which is Euclid's proposition 1 of book 2. However, there is very little else. A product ab of two lengths is not a length, so if c is a length then $ab + c$ does not make sense. Also, while abc is considered meaningful (it is a box with adjacent edges a, b, and c), $abcd$ is not, because the Greeks did not believe there could be mutually perpendicular lengths a, b, c, and d.

Another limitation is that finitely many additions and subtractions do *not* generally work for curved regions; for example, one would not expect to find a square equal in area to a circle by this method. More disappointing, the method does not generally work for polyhedra either. In particular, Dehn (1900) proved that a regular tetrahedron and a cube of the same volume are not "equal" by addition and subtraction of finitely many equal polyhedra.

Because of this, one is led to consider cutting polyhedra into *infinitely many* pieces. Euclid himself found the volume of a tetrahedron by cutting it into infinitely many prisms, as we will see in section 2.7.

2.6 NUMBER THEORY AND INDUCTION

In books 7–9 of the *Elements* Euclid develops something entirely different from the geometry in the first six books: it is what we would now call elementary number theory. His development looks superficially similar to the geometry—with geometric terminology such as "measures" instead of "divides," and careful step-by-step proofs—but no axioms are stated. This, perhaps, is because there were no doubts about the foundations of number theory as there were about the foundations of geometry. Nevertheless, we will see that Euclid was fleetingly aware of **induction**, which is recognized today as a fundamental principle—indeed, an *axiom*—of number theory.

Also similar to the first six books, Euclid's number theory propositions are a mixture of theorems and constructions. Proposition 1 of book 7, indeed, applies what we now call the **Euclidean algorithm** (section 1.4) to test whether two given numbers are relatively prime, that is, whether their greatest common divisor is 1. His proposition 2 shows, more generally, that the algorithm gives the **greatest common divisor** of any two numbers. In proposition 1 Euclid states the algorithm in its simplest form: "Two unequal numbers being set out, and the less being continually subtracted in turn from the greater. . . ." This is the form that

also applies to geometric quantities such as lengths and that will run forever if the lengths are in irrational ratio, as we saw in section 1.3.

In propositions 1 and 2 of book 7 Euclid assumes that the algorithm will terminate when the two quantities are natural numbers. Assuming also that any common divisor is preserved by subtraction, the greatest common divisor will survive (and be obvious) when the algorithm terminates with equal numbers. Both of these assumptions rest on induction.

1. Termination occurs because the algorithm produces smaller numbers, and *natural numbers cannot decrease forever*. This is the form of induction often called **infinite descent**.
2. The persistence of the greatest common divisor can be proved by the "base step, induction step" form of induction. Suppose the initial numbers a and b, with $a > b$, have a common divisor d, so

$$a = da' \quad \text{and} \quad b = db'.$$

Then the next pair $b = db'$ and $a - b = d(a' - b')$ also have the common divisor d. This is the base step. The induction step is similar. If the pair at step n, a_n and b_n, have common divisor d, then so have the pair at step $n + 1$, a_{n+1} and b_{n+1}, by the same argument as at step 1.

As we will see below, Euclid sometimes recognized when he was using infinite descent and pointed it out. But the "base step, induction step" idea does not occur in the *Elements*, if only because Euclid does not have notation (such as subscripts) to talk about sequences of arbitrary length. Instead of writing, say, $a_1, a_2, \ldots, a_n$ he would write a short sequence such as A, B, C and leave the reader to adapt the argument for the short sequence to one for a sequence of arbitrary length. This happens in his famous proof that there are infinitely many primes.

Primes

Euclid proved two famous theorems about prime numbers:

- *There are infinitely many primes*. Euclid actually proves a stronger result, which avoids mentioning infinity: given any finite collection of primes, we can (in principle) find another (book 9, proposition 20).

- *If a prime p divides a product ab of natural numbers a and b, then p divides a or p divides b* (book 7, proposition 30). We will call this the **prime divisor property**. It easily implies what we now call the **fundamental theorem of arithmetic**: each natural number > 1 has a prime factorization that is unique up to the order of factors.

To prove that there are infinitely many primes, Euclid needs the preliminary result that every natural number $k > 1$ has a prime divisor. This is proposition 31 in book 7 and is proved as follows. If k is not already prime, then k factorizes as ab, where $1 < a, b < k$. If one of a or b is prime, we are done. If not, continue splitting the factors into smaller factors. Since natural numbers cannot decrease forever (Euclid says it is "impossible in numbers"), eventually we find a prime factor of k.

Now the proof that there are infinitely many primes goes as follows. Suppose we are given some primes $p_1, p_2, \ldots, p_n$ (Euclid calls them just A, B, C). Consider the number

$$k = (p_1 p_2 \cdots p_n) + 1.$$

Then k is *not* divisible by any of $p_1, p_2, \ldots, p_n$. If p_i divides k, then p_i also divides $k - (p_1 p_2 \cdots p_n) = 1$, which is absurd. On the other hand, *some* prime p divides k, as we have just seen; hence p is a prime different from the given primes $p_1, p_2, \ldots, p_n$.

Comments. Unlike the variation of Euclid's proof often given today, his is *not* by contradiction. He does not suppose that there are finitely many primes and then look for a contradiction. Instead, he proves directly (and as finitely as possible) that there are infinitely many primes, by showing how to *increase* any given finite collection of primes. Also, strictly speaking, Euclid does not say take the product of the given primes and add 1. He actually says take the least common multiple of the given primes and add 1, but the least common multiple *is* the product in the case of distinct primes.

The second property, about a prime dividing a product, comes at the end of a rather lengthy sequence of consequences of the Euclidean algorithm. Today, using better notation and allowing negative integers, we can break the argument down to a shorter sequence of steps. We abbreviate "greatest common divisor" by gcd.