Undergraduate Texts in Mathematics

UTM

John Stillwell

Mathematics and Its History

A Concise Edition

Springer

John Stillwell
Department of Mathematics
University of San Francisco
San Francisco, CA, USA

ISSN 0172-6056 ISSN 2197-5604 (electronic)
Undergraduate Texts in Mathematics
ISBN 978-3-030-55192-6 ISBN 978-3-030-55193-3 (eBook)
https://doi.org/10.1007/978-3-030-55193-3

Mathematics Subject Classification: 00A05, 00A66, 01A05

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

1

The Theorem of Pythagoras

Preview

The Pythagorean theorem is the most appropriate starting point for a book on mathematics and its history. It is not only the oldest mathematical theorem, but also the source of three great streams of mathematical thought: numbers, geometry, and infinity.

The number stream begins with *Pythagorean triples*; triples of integers (a, b, c) such that $a^2 + b^2 = c^2$. The geometry stream begins with the interpretation of a^2, b^2, and c^2 as squares on the sides of a right-angled triangle with sides a, b, and hypotenuse c. The infinity stream begins with the discovery that $\sqrt{2}$, the hypotenuse of the right-angled triangle whose other sides are of length 1, is an *irrational* number.

These three streams are followed separately through Greek mathematics in Chapters 2, 3, and 4. The geometry stream resurfaces in Chapter 6, where it takes an *algebraic* turn. The basis of algebraic geometry is the possibility of describing points by numbers—their *coordinates*—and the bridge between coordinates and geometry is precisely the Pythagorean theorem, which defines length in terms of coordinates.

The Pythagorean theorem resurfaces in a new algebraic role in Chapter 16. Here it appears in the guise of the *inner product*, which introduces the concepts of length and angle into vector spaces.

J. Stillwell, *Mathematics and Its History*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-55193-3_1

1.1 Arithmetic and Geometry

If there is one theorem known to all mathematically educated people, it is surely the theorem of Pythagoras. It will be recalled as a property of right-angled triangles: the square of the hypotenuse equals the sum of the squares of the other two sides (Figure 1.1). The "sum" is of course the sum of areas and the area of a square of side l is l^2, which is why we call it "l squared." Thus the Pythagorean theorem can also be expressed by

$$a^2 + b^2 = c^2, \tag{1}$$

where a, b, c are the side lengths of the red triangle in Figure 1.1.

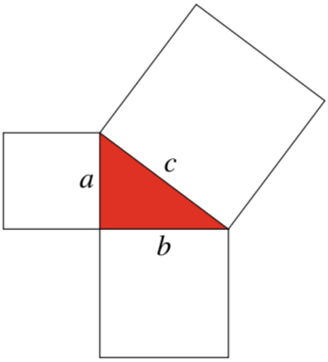

Figure 1.1: The Pythagorean theorem

Conversely, a solution of (1) by positive numbers a, b, c can be realized by a right-angled triangle with sides a, b and hypotenuse c. It is clear that we can draw perpendicular sides a, b for any given positive numbers a, b, and then the hypotenuse c must be a solution of (1) to satisfy the Pythagorean theorem. This converse view of the theorem becomes interesting when we notice that (1) has some very simple solutions. For example,

$$\begin{aligned} (a,b,c) &= (3,4,5), && (3^2 + 4^2 = 9 + 16 = 25 = 5^2), \\ (a,b,c) &= (5,12,13), && (5^2 + 12^2 = 25 + 144 = 169 = 13^2). \end{aligned}$$

It is thought that in ancient times such solutions may have been used for the construction of right angles. For example, by stretching a closed rope with 12 equally spaced knots one can obtain a $(3,4,5)$ triangle with right angle between the sides 3, 4, as seen in Figure 1.2.

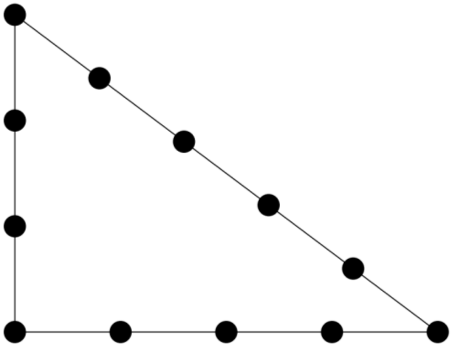

Figure 1.2: Right angle by rope stretching

Whether or not this is a practical method for constructing right angles, the very existence of a geometrical interpretation of a purely arithmetical fact like

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

is quite wonderful. At first sight, arithmetic and geometry seem to be completely unrelated realms. Arithmetic is based on counting, the epitome of a *discrete* (or *digital*) process. The facts of arithmetic can be clearly understood as outcomes of certain counting processes, and one does not expect them to have any meaning beyond this. Geometry, on the other hand, involves *continuous* rather than discrete objects, such as lines, curves, and surfaces. Continuous objects cannot be built from simple elements by discrete processes, and one expects to *see* geometrical facts rather than arrive at them by calculation.

The Pythagorean theorem was the first hint of a hidden, deeper relationship between arithmetic and geometry, and it has continued to hold a key position between these two realms throughout the history of mathematics. This has sometimes been a position of cooperation and sometimes one of conflict, as followed the discovery that $\sqrt{2}$ is irrational (see Section 1.5). It is often the case that new ideas emerge from such areas of tension, resolving the conflict and allowing previously irreconcilable ideas to interact fruitfully. The tension between arithmetic and geometry is, without doubt, the most profound in mathematics, and it has led to the most profound theorems. Since the Pythagorean theorem is the first of these, and the most influential, it is a fitting subject for our first chapter.

1.2 Pythagorean Triples

Pythagoras lived around 500 BCE, but the story of the Pythagorean theorem begins long before that, at least as far back as 1800 BCE in Babylonia. The evidence is a clay tablet, known as Plimpton 322, which systematically lists a large number of integer pairs (a, c) for which there is an integer b satisfying

$$a^2 + b^2 = c^2. \tag{1}$$

A translation of this tablet, together with its interpretation and historical background, was first published by Neugebauer and Sachs (1945). Integer triples (a, b, c) satisfying (1)—for example, $(3, 4, 5)$, $(5, 12, 13)$, $(8, 15, 17)$—are now known as *Pythagorean triples*. Presumably the Babylonians were interested in them because of their interpretation as sides of right-angled triangles, though this is not known for certain. At any rate, the problem of finding Pythagorean triples was considered interesting in other ancient civilizations that are known to have possessed the Pythagorean theorem; van der Waerden (1983) gives examples from China (between 200 BCE and 220 CE) and India (between 500 and 200 BCE). The most complete understanding of the problem in ancient times was achieved in Greek mathematics, between Euclid (around 300 BCE) and Diophantus (around 250 CE).

A general formula for generating Pythagorean triples is

$$a = (p^2 - q^2)r, \qquad b = 2qpr, \qquad c = (p^2 + q^2)r.$$

It is easy to see that $a^2 + b^2 = c^2$ when a, b, c are given by these formulas, and of course a, b, c will be integers if p, q, r are. Even though the Babylonians did not have the advantage of our algebraic notation, it is plausible that this formula, or the special case

$$a = p^2 - q^2, \qquad b = 2pq, \qquad c = p^2 + q^2$$

(which gives all solutions a, b, c, without common divisor and b even) was the basis for the triples they listed. Less general formulas have been attributed to Pythagoras himself (around 500 BCE) and Plato (see Heath (1921), Vol. 1, pp. 80–81); a solution equivalent to the general formula is given in Euclid's *Elements*, Book X (lemma following Prop. 28). As far as we know, this is the first statement of the general solution and the first proof that it is general. Euclid's proof is essentially arithmetical, as one would expect since the problem seems to belong to arithmetic.

However, there is a far more striking solution, which uses the geometric interpretation of Pythagorean triples. This emerges from the work of Diophantus, and it is described in the next section.

Exercises

The integer pairs (a, c) in Plimpton 322 are shown in Figure 1.3.

a	c
119	169
3367	4825
4601	6649
12709	18541
65	97
319	481
2291	3541
799	1249
481	769
4961	8161
45	75
1679	2929
161	289
1771	3229
56	106

Figure 1.3: Pairs in Plimpton 322

1.2.1 For each pair (a, c) in the table, compute $c^2 - a^2$, and confirm that it is a perfect square, b^2. (Computer assistance is recommended.)

You should notice that in most cases b is a "rounder" number than a or c.

1.2.2 Show that most of the numbers b are divisible by 60, and that the rest are divisible by 30 or 12.

Such numbers were in fact exceptionally "round" for the Babylonians, because 60 was the base for their system of numerals. It looks like they computed Pythagorean triples starting with the "round" numbers b and that the column of b values later broke off the tablet.

Euclid's formula for Pythagorean triples comes out of his theory of divisibility, which we take up in Section 3.3. Divisibility is also involved in some basic properties of Pythagorean triples, such as their evenness or oddness.

1.2.3 Show that any integer square leaves remainder 0 or 1 on division by 4.

1.2.4 Deduce from Exercise 1.2.3 that if (a, b, c) is a Pythagorean triple then a and b cannot both be odd.

1.3 Rational Points on the Circle

We know from Section 1.1 that a Pythagorean triple (a, b, c) can be realized by a triangle with sides a, b and hypotenuse c. This in turn yields a triangle with fractional (or *rational*) number sides $x = a/c$, $y = b/c$ and hypotenuse 1. All such triangles can be fitted inside the circle of radius 1 as shown in Figure 1.4. The sides x and y become what we now call the *coordinates* of the point P on the circle. The Greeks did not use this language, but they could derive the relationship between x and y we call the *equation of the circle*. Since

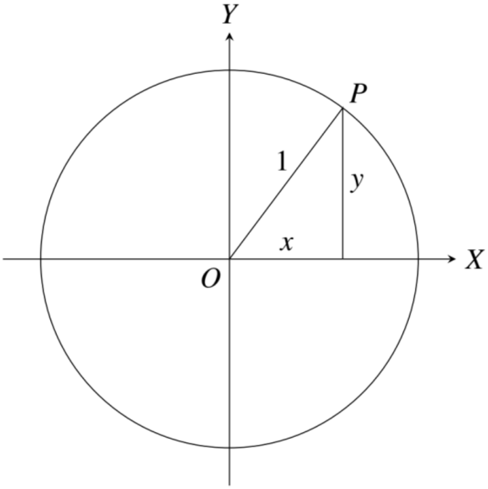

Figure 1.4: The unit circle

$$a^2 + b^2 = c^2 \tag{1}$$

we have

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1,$$

so the relationship between $x = a/c$ and $y = b/c$ is

$$x^2 + y^2 = 1. \tag{2}$$

Consequently, finding integer solutions of (1) is equivalent to finding rational solutions of (2), or finding *rational points* on the curve (2).

Such problems are now called *Diophantine*, after Diophantus, who was the first to deal with them seriously and successfully. *Diophantine equations* have acquired the more special connotation of equations for which integer solutions are sought, although Diophantus himself sought only rational solutions. (There is an interesting open problem that turns on this distinction. Matiyasevich (1970) proved that there is no algorithm for deciding which polynomial equations have integer solutions. It is not known whether there is an algorithm for deciding which polynomial equations have *rational* solutions.)

Most of the problems solved by Diophantus involve quadratic or cubic equations, usually with one obvious trivial solution. Diophantus used the obvious solution as a stepping stone to the nonobvious, but no account of his method survived. It was ultimately reconstructed by Fermat and Newton in the 17th century, and this *chord and tangent construction* will be considered later. Here, we need it only for the equation $x^2 + y^2 = 1$, which is an ideal showcase for the method in its simplest form (chord only).

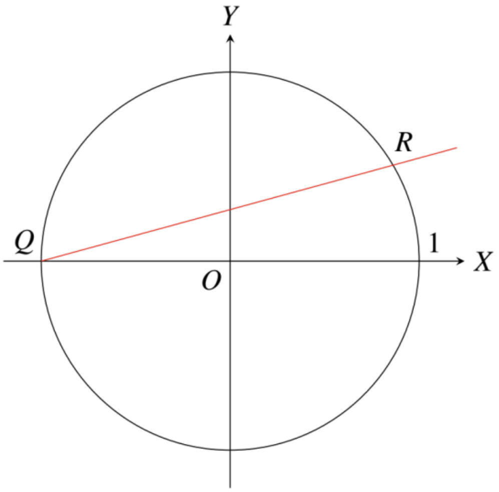

Figure 1.5: Construction of rational points

A trivial solution of this equation is $x = -1, y = 0$, which is the point Q on the unit circle (Figure 1.5). After a moment's thought, one realizes that a line through Q, with rational slope t,

$$y = t(x + 1) \tag{3}$$

will meet the circle at a second rational point R. This is because substitution of $y = t(x+1)$ in $x^2 + y^2 = 1$ gives a quadratic equation with rational coefficients and one rational solution ($x = -1$); hence the second solution must also be a rational value of x. But then the y value of this point will also be rational, since t and x will be rational in (3). Conversely, the chord joining Q to any other rational point R on the circle will have a rational slope. Thus by letting t run through all rational values, we find all rational points $R \neq Q$ on the unit circle.

What are these points? We find them by solving the equations just discussed. Substituting $y = t(x+1)$ in $x^2 + y^2 = 1$ gives

$$x^2 + t^2(x+1)^2 = 1,$$

or

$$x^2(1+t^2) + 2t^2x + (t^2 - 1) = 0.$$

This quadratic equation in x has solutions -1 and $(1-t^2)/(1+t^2)$. The nontrivial solution $x = (1-t^2)/(1+t^2)$, when substituted in (3), gives $y = 2t/(1+t^2)$.

Exercises

The parameter t in the pair $\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)$ runs through all rational numbers if $t = q/p$ and p, q run through all pairs of integers.

1.3.1 Deduce that if (a, b, c) is any Pythagorean triple then

$$\frac{a}{c} = \frac{p^2 - q^2}{p^2 + q^2}, \quad \frac{b}{c} = \frac{2pq}{p^2 + q^2}$$

for some integers p and q.

1.3.2 Use Exercise 1.3.1 to prove Euclid's formula for Pythagorean triples, assuming b even. (Remember, a and b are not both odd.)

The triples (a, b, c) in Plimpton 322 seem to have been computed to provide right-angled triangles covering a range of shapes—their angles actually follow a decreasing sequence in roughly equal steps. Figure 1.6 shows the lines with slope a/b, ranging from the top value $119/120$ for the top line in Plimpton 322, to 56/90 for the bottom line.

This raises the question, can the shape of any right-angled triangle be approximated by a Pythagorean triple?

1.3.3 Show that any right-angled triangle with hypotenuse 1 may be approximated arbitrarily closely by one with rational sides.

b	a	c	a/b
120	119	169	0.9917
3456	3367	4825	0.9742
4800	4601	6649	0.9585
13500	12709	18541	0.9414
72	65	97	0.9028
360	319	481	0.8861
2700	2291	3541	0.8485
960	799	1249	0.8323
600	481	769	0.8017
6480	4961	8161	0.7656
60	45	75	0.7500
2400	1679	2929	0.6996
240	161	289	0.6708
2700	1771	3229	0.6559
90	56	106	0.6222

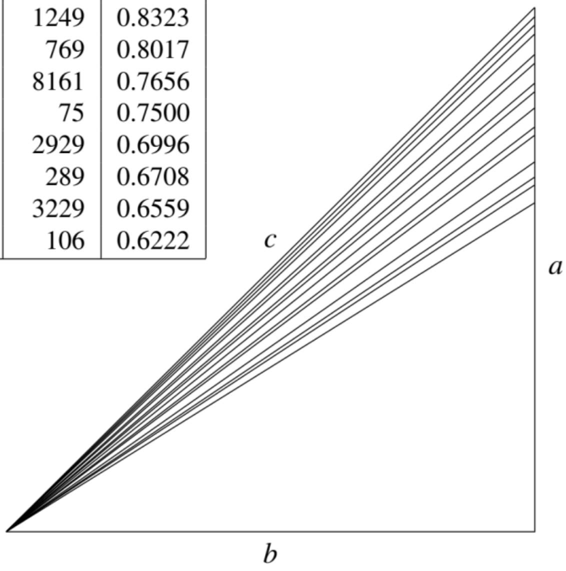

Figure 1.6: Lines of slope a/b corresponding to entries in Plimpton 322

Some important trigonometry may be gleaned from Diophantus's method if we compare the angle at O in Figure 1.4 with the angle at Q in Figure 1.5. The two angles are shown in Figure 1.7, and high school geometry shows that the angle at Q is half the angle at O.

1.3.4 Why does the angle at Q equal $\theta/2$? (Hint: consider angles in the red triangle.)

1.3.5 Use Figure 1.7 to show that $t = \tan\frac{\theta}{2}$ and

$$\cos\theta = \frac{1-t^2}{1+t^2}, \quad \sin\theta = \frac{2t}{1+t^2}.$$

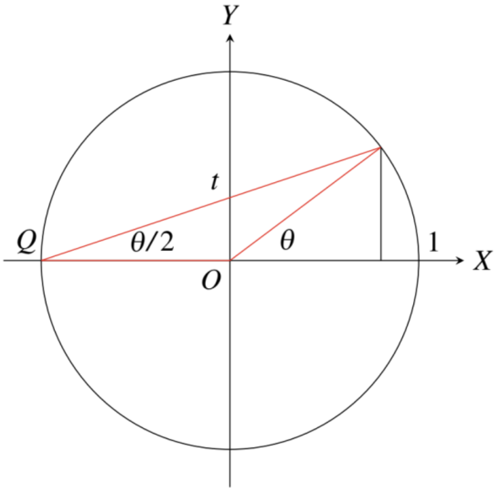

Figure 1.7: Angles in a circle

1.4 Right-Angled Triangles

It is high time we looked at the Pythagorean theorem from the traditional point of view, as a theorem about right-angled triangles; however, we will be rather brief about its proof. It is not known how the theorem was first proved, but probably it was by simple manipulations of area, perhaps suggested by rearrangement of floor tiles. Just how easy it can be to prove the Pythagorean theorem is shown by Figure 1.8, given by Heath (1925) in his edition of Euclid's *Elements*, Vol. 1, p. 354. Each large square contains four copies of the given right-angled triangle. Subtracting these four triangles from the large square leaves, on the one hand (Figure 1.8, *right*), the sum of the squares on the two sides of the triangle. On the other hand (*left*), it also leaves the square on the hypotenuse. This proof, like the hundreds of others that have been given for the Pythagorean theorem, rests on certain geometric assumptions. It is in fact possible to transcend geometric assumptions by using numbers as the foundation for geometry, and the Pythagorean theorem then becomes true almost by definition, as an immediate consequence of the definition of distance (see Section 1.5).

To the Greeks, however, it did not seem possible to build geometry on the basis of numbers, due to a conflict between their notions of number and length. In the next section we will see how this conflict arose.

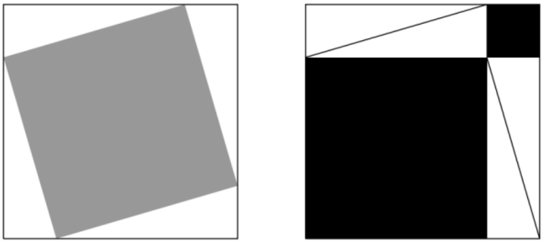

Figure 1.8: Proof of the Pythagorean theorem

Exercises

A way to see the Pythagorean theorem in a tiled floor was suggested by Magnus (1974), p. 159, and it is shown in Figure 1.9. (The dotted squares are not tiles; they are a hint.)

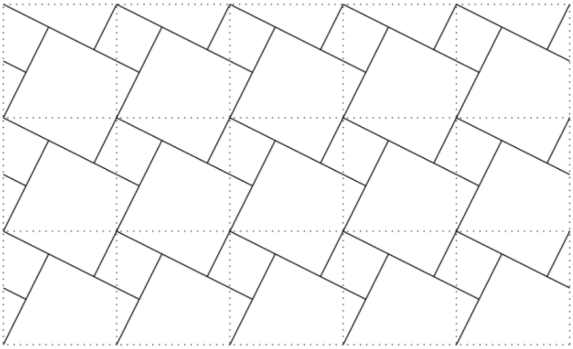

Figure 1.9: Pythagorean theorem in a tiled floor

1.4.1 What has this figure to do with the Pythagorean theorem?

Euclid's first proof of the Pythagorean theorem, in Book I of the *Elements*, is also based on area. It depends only on the fact that triangles with the same base and height have equal area, though it involves a rather complicated figure. In Book VI, Proposition 31, he gives another proof, based on similar triangles (Figure 1.10).

1.4.2 Show that the three triangles in Figure 1.10 are similar, and hence prove the Pythagorean theorem by equating ratios of corresponding sides.

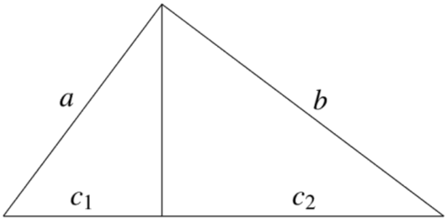

Figure 1.10: Another proof of the Pythagorean theorem

1.5 Irrational Numbers

We have mentioned that the Babylonians, although probably aware of the geometric meaning of the Pythagorean theorem, devoted most of their attention to the whole-number triples it had brought to light, the Pythagorean triples. Pythagoras and his followers were even more devoted to whole numbers. It was they who discovered the role of numbers in musical harmony: dividing a vibrating string in two raises its pitch by an octave, dividing in three raises the pitch another fifth, and so on. This great discovery, the first clue that the physical world might have an underlying mathematical structure, inspired them to seek numerical patterns, which to them meant *whole-number* patterns, everywhere. Imagine their consternation when they found that the Pythagorean theorem led to quantities that were not numerically computable. They found lengths that were *incommensurable*, that is, not measurable as integer multiples of the same unit. The ratio between such lengths is therefore not a ratio of whole numbers, hence in the Greek view not a ratio at all, or *irrational*.

The incommensurable lengths discovered by the Pythagoreans were the side and diagonal of the unit square. It follows immediately from the Pythagorean theorem that

$$(\text{diagonal})^2 = 1 + 1 = 2.$$

Hence if the diagonal and side are in the ratio m/n (where m and n can be assumed to have no common divisor), we have

$$m^2/n^2 = 2,$$

whence

$$m^2 = 2n^2.$$

The Pythagoreans were interested in odd and even numbers, so they probably observed that the latter equation, which says that m^2 is even, also implies that m is even, say $m = 2p$. But if

$$m = 2p,$$

then

$$2n^2 = m^2 = 4p^2;$$

hence

$$n^2 = 2p^2,$$

which similarly implies that n is even, contrary to the hypothesis that m and n have no common divisor. (This proof is in Aristotle's *Prior Analytics*. An alternative, more geometric, proof is mentioned in Section 3.4.)

This discovery had profound consequences. Legend has it that the first Pythagorean to make the result public was drowned at sea (see Heath (1921), Vol. 1, pp. 65, 154). It led to a split between the theories of number and space that was not healed until the 19th century (if then, some believe). The Pythagoreans could not accept $\sqrt{2}$ as a number, but no one could deny that it was the diagonal of the unit square. Consequently, geometrical quantities had to be treated separately from numbers or, rather, without mentioning any numbers except rationals. Greek geometers thus developed ingenious techniques for precise handling of arbitrary lengths in terms of rationals, known as the *theory of proportions* and the *method of exhaustion*.

As we will see in Chapter 4, these techniques made necessary use of *infinity*—something that the Greeks were very reluctant to do.

The Reconciliation of Numbers with Geometry

As we now know, it is not necessary to deny that $\sqrt{2}$ is a number, or to do geometry without applying the processes of arithmetic to lengths, areas, and volumes. In the 1620s, Fermat and Descartes realized that, if lengths are viewed as numbers, then each point P in the plane is given by an ordered pair (x, y) of numbers, called the *coordinates* of P. The coordinates x and y are respectively the horizontal and vertical distances of P from an origin O. We tell the story of their discovery, and the reasons for its success, in Chapter 6.

In coordinate geometry one can *define* the distance between any two points, guided by none other than the Pythagorean theorem. If $P_1 = (x_1, y_1)$

and $P_2 = (x_2, y_2)$ then the line P_1P_2 from P_1 to P_2 is the hypotenuse of a triangle with horizontal side $x_2 - x_1$ and vertical side $y_2 - y_1$ (Figure 1.11).

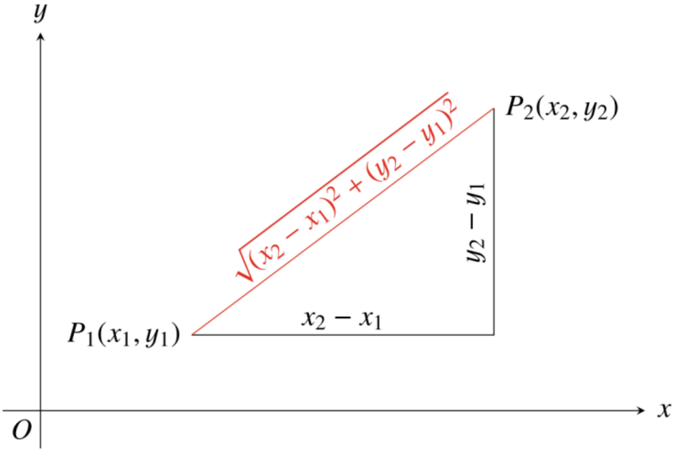

Figure 1.11: Distance via the Pythagorean theorem

Since the square of the hypotenuse is the sum of the squares on the other two sides,

$$(x_2 - x_1)^2 + (y_2 - y_1)^2,$$

we should define

$$\text{length of } P_1P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

It follows, for example, that the points (x, y) at distance 1 from O satisfy the equation $x^2 + y^2 = 1$, which we called the *equation of the (unit) circle* in Section 1.3. The coordinate geometry of Fermat and Descartes is part of what is now called *algebraic geometry*, a vast expansion of Greek geometry. Algebraic geometry was made possible by 16th century discoveries in algebra, which brought the study of curves into alignment with the study of polynomial equations.

A coordinate geometry closer in content to Greek geometry, particularly that of Euclid, was developed by Grassmann in the 1840s. Grassmann's geometry is part of what we now call *linear algebra*, and its key concept—the *inner product*—is also inspired by the Pythagorean theorem. For more on linear algebra and the inner product, see Section 16.2.

Exercises

The crucial step in the proof that $\sqrt{2}$ is irrational is showing that m^2 even implies m is even or, equivalently, that m odd implies m^2 odd. It is worth taking a closer look at why this is true.

1.5.1 Writing an arbitrary odd number m in the form $2q + 1$, for some integer q, show that m^2 also has the form $2r + 1$, which shows that m^2 is also odd.

You probably did some algebra like this in Exercise 1.2.3, but if not, here is your chance:

1.5.2 Show that the square of $2q + 1$ is in fact of the form $4s + 1$, and hence explain why every integer square leaves remainder 0 or 1 on division by 4.

2

Greek Geometry

Preview

Geometry was the first branch of mathematics to become highly developed. The concepts of “theorem” and “proof” originated in geometry, and most mathematicians until recent times were introduced to their subject through the geometry in Euclid’s *Elements*.

In the *Elements* one finds the first system for deriving theorems from supposedly self-evident statements called *axioms*. Euclid’s axioms are incomplete and one of them, the so-called *parallel* axiom, is not as obvious as the others. Nevertheless, it took over 2000 years to produce a clearer foundation for geometry.

The climax of the *Elements* is the investigation of the regular polyhedra, five symmetric figures in three-dimensional space. The five regular polyhedra make several appearances in mathematical history, most importantly in the theory of symmetry—*group theory*—discussed in Chapter 14.

The *Elements* contains not only proofs but also many *constructions*, by ruler and compass. However, three constructions are conspicuous by their absence: duplication of the cube, trisection of the angle, and squaring the circle. These problems were not properly understood until the 19th century, when they were resolved (in the negative) by algebra and analysis.

The only curves in the *Elements* are circles, but the Greeks studied many other curves, such as the conic sections. Again, many problems that the Greeks could not solve were later clarified by algebra. In particular, curves can be classified by *degree*, and the conic sections are the curves of degree 2, as we will see in Chapter 6.

J. Stillwell, *Mathematics and Its History*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-55193-3_2

2.1 The Deductive Method

> 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 about the philosopher Thomas Hobbes (1588–1679), from Aubrey's *Brief Lives*, beautifully captures the force of Greece's most important contribution to mathematics, the deductive method. (The proposition mentioned, incidentally, is the Pythagorean theorem.)

We have seen that significant results were *known* before the period of classical Greece, but the Greeks were the first to find results by deduction from previously established results, resting ultimately on the most evident possible statements, called *axioms*. Thales (624–547 BCE) is thought to be the originator of this method (see Heath (1921), p. 128), and by 300 BCE Euclid's *Elements* set the standard for mathematical rigor until the 19th century. But the *Elements* is difficult, so in time it was boiled down to the simplest and driest propositions about lines, angles, and circles. These propositions are based on the following axioms (in the translation of Heath (1925), p. 154), which Euclid called *postulates* and *common notions*.

Postulates

Let the following be postulated:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance.
4. That all right angles are equal to one another.
5. 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.

Common Notions

1. Things which are equal to the same thing are also equal to one another.
2. If equals be added to equals, the wholes are equal.
3. If equals be subtracted from equals, the remainders are equal.
4. Things which coincide with one another are equal to one another.
5. The whole is greater than the part.

It appears that Euclid's intention was to deduce geometric propositions from visually evident statements (the postulates) using evident principles of logic (the common notions). Actually, he often made unconscious use of visually plausible assumptions that are not among his postulates. His very first proposition used the unstated assumption that two circles meet if the center of each is on the circumference of the other (Heath (1925), p. 242). Nevertheless, such flaws were not noticed until the 19th century, and they were rectified by Hilbert (1899). By themselves, they probably would not have been enough to end the *Elements*' run of 22 centuries as a leading textbook. The *Elements* was overthrown by more serious mathematical upheavals in the 19th century. The so-called non-Euclidean geometries, using alternatives to Euclid's fifth postulate (the *parallel axiom*), developed to the point where the old axioms could no longer be considered self-evident (see Chapter 13). At the same time, the concept of number matured to the point where irrational numbers became acceptable, and indeed preferable to intuitive geometric concepts, in view of the doubts about what the self-evident truths of geometry really were.

The outcome was a more adaptable language for geometry in which "points," "lines," and so on, could be defined, usually in terms of numbers, so as to suit the type of geometry under investigation. Such a development was long overdue. Even in Euclid's time the Greeks were investigating curves more complicated than circles, which did not fit conveniently in Euclid's system. Descartes (1637) introduced the coordinate method, which gives a single framework for handling both Euclid's geometry and higher curves (see Chapter 6), but it was not at first realized that coordinates allowed geometry to be entirely rebuilt on numerical foundations.

The comparatively trivial step (for us) of passing to axioms about numbers from axioms about points had to wait until the 19th century, when geometric axioms about points lost authority and number-theoretic axioms gained it. We say about these developments later (and of problems with the authority of axioms in general, which arose in the 20th century). For the remainder of this chapter we will look at some important nonelementary topics in Greek geometry, using the coordinate framework where convenient.

Exercises

Euclid's Common Notions 1 and 4 define what we now call an *equivalence relation*, which is not necessarily the equality relation. In fact, the kind of relation Euclid had in mind was equality in *some* geometric quantity such as length or angle (but not necessarily equality in all respects—the latter is what he meant by "coinciding"). An equivalence relation $\cong$ is normally defined by three properties. For any a, b and c:

$$a \cong a, \tag{reflexive}$$

$$a \cong b \;\Rightarrow\; b \cong a, \tag{symmetric}$$

$$a \cong b \text{ and } b \cong c \;\Rightarrow\; a \cong c. \tag{transitive}$$

2.1.1 Explain how Common Notions 1 and 4 may be interpreted as the transitive and reflexive properties. Note that the natural way to write Common Notion 1 symbolically is slightly different from the statement of transitivity above.

2.1.2 Show that the symmetric property follows from Euclid's Common Notions 1 and 4.

Hilbert (1899) took advantage of Euclid's Common Notions 1 and 4 in his rectification of Euclid's axiom system. He *defined* equality of length by postulating a transitive and reflexive relation on line segments, and stated transitivity in the style of Euclid, so that the symmetric property was a consequence.

2.2 The Regular Polyhedra

Greek geometry is virtually complete as far as the elementary properties of plane figures are concerned. It is fair to say that only a handful of interesting elementary propositions about triangles and circles have been discovered since Euclid's time. Solid geometry is much more challenging, even today, so it is understandable that it was left in a less complete state by the Greeks. Nevertheless, they made some very impressive discoveries and managed to complete one of the most beautiful chapters in solid geometry, the enumeration of the regular polyhedra. The five possible regular polyhedra are shown in Figure 2.1. (Images courtesy of Wikimedia.)

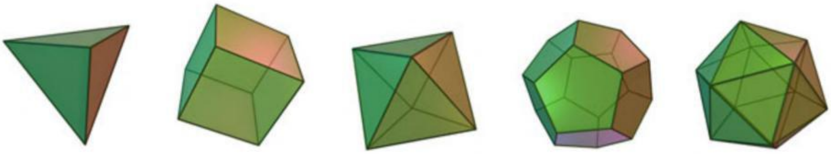

Figure 2.1: Tetrahedron, cube, octahedron, dodecahedron, icosahedron

Each polyhedron is convex and is bounded by a number of congruent polygonal faces, the same number of faces meet at each vertex, and in each face all the sides and angles are equal, hence the term *regular polyhedron*. A regular polyhedron is a spatial figure analogous to a regular polygon in the plane. But whereas there are regular polygons with any number $n \geq 3$ of sides, there are only five regular polyhedra.

This fact is easily proved and may go back to the Pythagoreans (see, for example Heath (1921), p. 159). One considers the possible polygons that can occur as faces, their angles, and the numbers of them that can occur at a vertex. For a 3-gon (triangle) the angle is $\pi/3$, so three, four, or five can occur at a vertex, but six cannot, as this would give a total angle 2π and the vertex would be flat. For a 4-gon the angle is $\pi/2$, so three can occur at a vertex, but not four. For a 5-gon the angle is $3\pi/5$, so three can occur at a vertex, but not four. For a 6-gon the angle is $2\pi/3$, so not even three can occur at a vertex. But at least three faces must meet at each vertex of a polyhedron, so 6-gons (and, similarly, 7-gons, 8-gons, . . .) cannot occur as faces of a regular polyhedron. This leaves only the five possibilities just listed, which correspond to the five known regular polyhedra.

But do these five really exist? There is no trouble constructing the tetrahedron, cube, or octahedron, but it is not clear that, say, 20 equilateral triangles will fit together to form a closed surface. Euclid found this problem difficult enough to be placed near the end of the *Elements*, and few of his readers ever mastered his solution. A beautiful direct construction was given by Luca Pacioli, a friend of Leonardo da Vinci's, in his book *De divina proportione* (1509). Pacioli's construction uses three copies of the *golden rectangle*, with sides 1 and $(1 + \sqrt{5})/2$, interlocking as in Figure 2.2. The 12 vertices define 20 triangles such as ABC, and it suffices to show that these are equilateral, that is, $AB = 1$. This is a straightforward exercise in the Pythagorean theorem (Exercise 2.2.2).

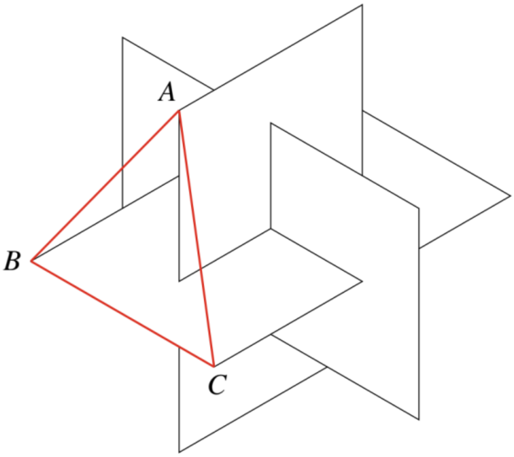

Figure 2.2: Pacioli's construction of the icosahedron

The regular polyhedra will make another important appearance in yet another 19th-century development, the theory of finite groups and Galois theory. See Chapter 14. Before the regular polyhedra made this triumphant comeback, they also took part in a famous fiasco: the Kepler (1596) theory of planetary distances. Kepler's theory is summarized by his famous diagram (Figure 2.3) of the five polyhedra, nested in such a way as to produce six spheres of radii proportional to the distances of the six planets then known. Unfortunately, although mathematics could not permit any more regular polyhedra, nature could permit more planets, and Kepler's theory was ruined when Uranus was discovered in 1781.

Exercises

The ratios between successive radii in Kepler's construction depend on what may be called the *inradius* and *circumradius* of each polyhedron—the radii of the spheres that touch it on the inside and the outside. It happens that the ratio

$$\frac{\text{circumradius}}{\text{inradius}}$$

is the same for the cube and the octahedron, and it is also the same for the dodecahedron and the icosahedron. This implies that the cube and octahedron can be exchanged in Kepler's construction, as can the dodecahedron and the icosahedron. Thus there are at least four different arrangements of the regular polyhedra that yield the same sequence of radii.

It is easy to see why the cube and the octahedron are interchangeable.

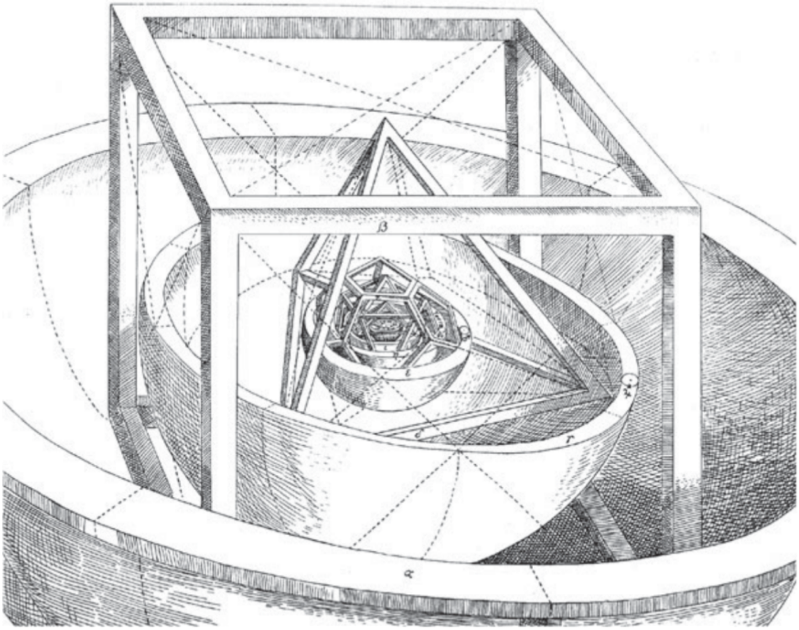

Figure 2.3: Kepler's diagram of the polyhedra

2.2.1 Show that $\frac{\text{circumradius}}{\text{inradius}} = \sqrt{3}$ for both the cube and the octahedron.

To compute circumradius/inradius for the icosahedron and the dodecahedron is quite difficult, and we will not pursue it further, other than verifying that Pacioli's construction gives a figure bounded by equilateral triangles.

2.2.2 Check Pacioli's construction: use the Pythagorean theorem to show that $AB = BC = CA$ in Figure 2.2. (It may help to use the additional fact that $\tau = (1 + \sqrt{5})/2$ satisfies $\tau^2 = \tau + 1$.)

2.3 Ruler and Compass Constructions

Greek geometers prided themselves on their logical purity; nevertheless, they were guided by intuition about physical space. One aspect of Greek geometry that was peculiarly influenced by physical considerations was the theory of constructions. Much of the elementary geometry of straight lines and circles can be viewed as the theory of constructions by ruler and compass. (By a "ruler" we mean simply a straightedge; it is not assumed to have any marks on it.) The very subject matter, lines and circles, reflects the instruments used to draw them. And many of the elementary problems of geometry—for example, to bisect a line segment or angle, construct a perpendicular, or draw a circle through three given points—can be solved by ruler and compass constructions.

When coordinates are introduced, it is not hard to show that the points constructible from points $P_1, \ldots, P_n$ have coordinates in the set of numbers generated from the coordinates of $P_1, \ldots, P_n$ by the operations $+$, $-$, $\times$, $\div$, and $\sqrt{\ }$ (see Moise (1963) or the exercises to Section 5.3). Square roots arise, of course, because of the Pythagorean theorem: if points (a, b) and (c, d) have been constructed, then so has the distance $\sqrt{(c-a)^2 + (d-b)^2}$ between them (Section 1.5). Conversely, it is possible to construct $\sqrt{l}$ for any given length l (Exercise 2.3.2).

Seen from this viewpoint, ruler and compass constructions look very special and unlikely to yield numbers such as $\sqrt[3]{2}$, for example. Just this number comes up in the classical Greek problem called *duplication of the cube*, since doubling the volume of a cube amounts to multiplying its side $\sqrt[3]{2}$. Other notorious problems were *trisection of the angle* and *squaring the circle*.[1] The latter problem was to construct a square equal in area to a given circle or to construct the number π, which amounts to the same thing. They sought ruler and compass solutions, though the possibility of a negative solution was admitted and solutions by less elementary means were tolerated. We will see some of these in the next sections.

The impossibility of solving these problems by ruler and compass constructions was not proved until the 19th century. For the duplication of the cube and trisection of the angle, impossibility was shown by Wantzel (1837). Wantzel seldom receives credit for settling these problems, which had baffled the best mathematicians for 2000 years, perhaps because his methods were superseded by the more powerful theory of algebraic numbers (see Chapter 16).

The impossibility of squaring the circle was proved by Lindemann (1882), in a very strong way. Not only is π undefinable by rational operations and square roots; it is also *transcendental*, that is, not the root of any polynomial equation with rational coefficients. Like Wantzel's work, this was a rare example of a major result proved by a minor mathematician. In

[1] The term "squaring," or its Latin equivalent "quadrature," later became a general term for finding the area of curved regions, particularly in the 17th century, when calculus solved many such problems. Since ancient times the "squaring the circle" has been a popular phrase for trying to do the impossible.

Lindemann's case the explanation is perhaps that a major step had already been taken when Hermite (1873) proved the transcendence of e. Accessible proofs of both these results can be found in Klein (1924). Lindemann's subsequent career was mathematically undistinguished, even embarrassing. In response to skeptics who thought his success with π had been a fluke, he took aim at the most famous unsolved problem in mathematics, "Fermat's last theorem" (see Chapter 10 for the origin of this problem). His efforts fizzled out in a series of inconclusive papers, each one correcting an error in the one before. Fritsch (1984) has written an interesting biographical article on Lindemann.

One ruler and compass problem is still open: which regular n-gons are constructible? Gauss discovered in 1796 that the 17-gon is constructible and then showed that a regular n-gon is constructible if and only if $n = 2^m p_1 p_2 \cdots p_k$, where the p_i are distinct primes of the form $2^{2^h} + 1$. (This problem is also known as *circle division*, because it is equivalent to dividing the circumference of a circle, or the angle 2π, into n equal parts.) The proof of necessity was actually completed by Wantzel (1837). However, it is still not explicitly known what these primes are, or even whether there are infinitely many of them. The only ones known are for $h = 0, 1, 2, 3, 4$.

Exercises

Many of the constructions made by the Greeks are simplified by translating them into algebra, where it turns out that constructible lengths are those that can be built from known lengths by the operations of $+$, $-$, $\times$, $\div$, and $\sqrt{\ }$. It is therefore enough to know constructions for these five basic operations. Addition and subtraction are obvious, and the other operations are covered in the following exercises, together with an example in which algebra is a distinct advantage.

2.3.1 Show, using similar triangles, that if lengths l_1 and l_2 are constructible, then so are $l_1 l_2$ and l_1/l_2.

2.3.2 Use similar triangles to explain why $\sqrt{l}$ is the length shown in Figure 2.4, and hence show that $\sqrt{l}$ is constructible from l.

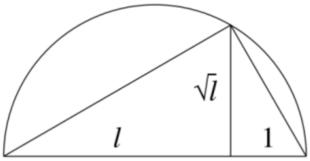

Figure 2.4: Square root construction

One of the finest ruler and compass constructions from ancient times is that of the regular pentagon, which includes, yet again, the golden ratio $\tau = (1 + \sqrt{5})/2$. Knowing (from the questions above) that this number is constructible, it becomes easy for us to construct the pentagon itself.

2.3.3 By finding some parallels and similar triangles in Figure 2.5, show that the diagonal x of the regular pentagon of side 1 satisfies $x/1 = 1/(x-1)$.

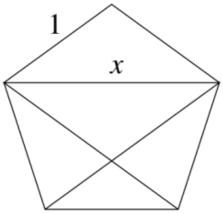

Figure 2.5: The regular pentagon

2.3.4 Deduce from Exercise 2.3.3 that the diagonal of the pentagon is $(1+\sqrt{5})/2$ and hence that the regular pentagon is constructible.

2.4 Conic Sections

Conic sections are the curves obtained by cutting a circular cone by a plane: ellipses (including circles), parabolas, and hyperbolas (Figure 2.6, left to right). Today we know the conic sections better by their equations:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad \text{(ellipse)}$$

$$y = ax^2, \qquad \text{(parabola)}$$

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \qquad \text{(hyperbola)}$$

More generally, any second-degree equation represents a conic section or a pair of straight lines, a result that was proved by Descartes (1637).

The names “ellipse,” “parabola”, and “hyperbola” come from the Greek, meaning roughly “too little,” “alongside,” and “too much.” The ellipse arises by cutting with a plane that slopes too little (to make an infinite curve), the parabola from a plane parallel to one side of the cone, and the hyperbola from a plane that slopes too much to avoid hitting the other part of the cone, so it produces a curve with two branches.

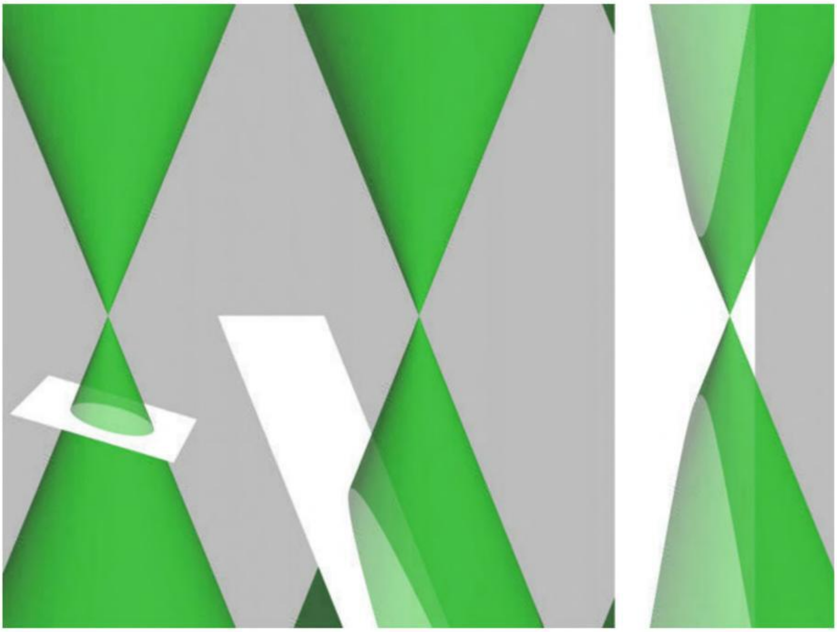

Figure 2.6: Ellipse, parabola, hyperbola

The invention of conic sections is attributed to Menaechmus (fourth century BCE), a contemporary of Alexander the Great. Alexander is said to have asked Menaechmus for a crash course in geometry, but Menaechmus refused, saying, "There is no royal road to geometry." Menaechmus used conic sections to give a very simple solution to the problem of duplicating the cube. In algebraic notation, this can be described as finding the intersection of the parabola $y = \frac{1}{2}x^2$ with the hyperbola $xy = 1$. This yields

$$x\frac{1}{2}x^2 = 1 \qquad \text{or} \qquad x^3 = 2.$$

The theory and practice of conic sections finally came together when Kepler (1609) found the orbits of the planets to be ellipses, and Newton (1687) explained this fact by his law of gravitation. This wonderful vindication of the theory of conic sections has often been seen as basic research receiving its long overdue reward, but perhaps one can also see it as a rebuke to Greek disdain for applications. As for Kepler himself ... to the end of his days he was proudest of his theory explaining the distances of the planets in terms of the five regular polyhedra (Section 2.2).

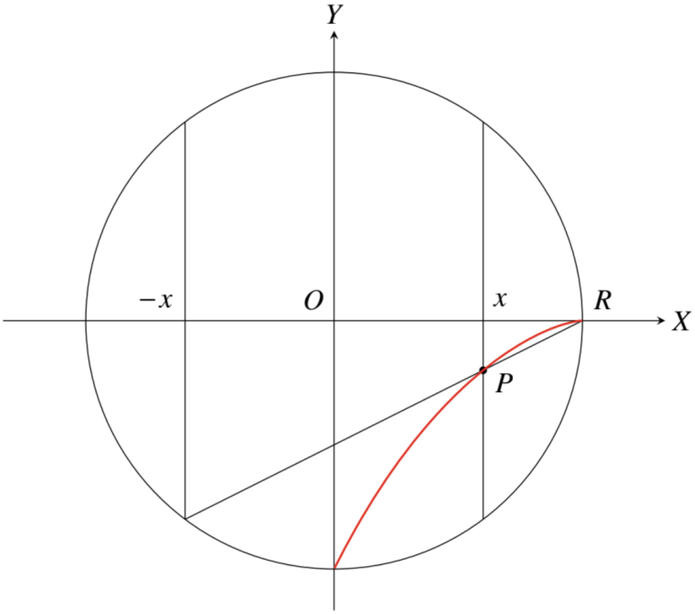

Figure 2.8: Construction of the cissoid

The Spiric Sections of Perseus (around 150 BCE)

Apart from the sphere, cylinder, and cone—whose sections are all conic sections—one of the few surfaces studied by the Greeks was the *torus*. This surface, generated by rotating a circle about an axis outside the circle, but in the same plane, was called a *spira* by the Greeks—hence the name spiric sections for the sections by planes parallel to the axis. These sections, which were first studied by Perseus, have four qualitatively distinct forms (see Figure 2.9).

These forms—convex ovals, “squeezed” ovals, the figure 8, and pairs of ovals—were rediscovered in the 17th century when analytic geometers looked at curves of degree 4, of which the spiric sections are examples. For suitable choice of torus, the figure 8 curve becomes the *lemniscate of Bernoulli* and the convex ovals become *Cassini ovals*. Cassini (1625–1712) was a distinguished astronomer but an opponent of Newton’s theory of gravitation. He rejected Kepler’s ellipses and instead proposed Cassini ovals as orbits for the planets.

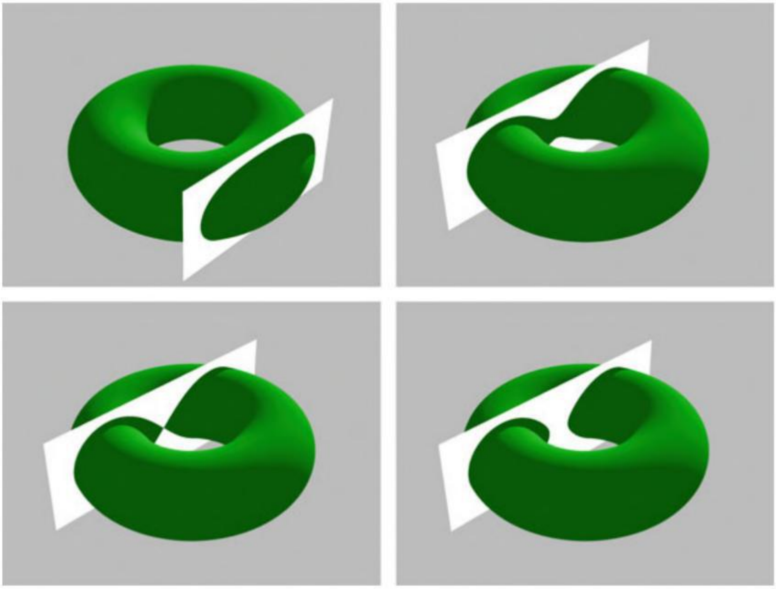

Figure 2.9: Spiric sections

The Epicycles of Ptolemy (140 CE)

These curves are known from a famous astronomical work, the *Almagest* of Claudius Ptolemy. Ptolemy himself attributes the idea to Apollonius. It seems almost certain that this is the Apollonius who mastered conic sections, which is ironic, because epicycles were his candidates for the planetary orbits, destined to be defeated by those very same conic sections.

An epicycle, in its simplest form, is the path traced by a point on a circle that rolls on another circle (Figure 2.10). More complicated epicycles can be defined by having a third circle roll on the second, and so on. The Greeks introduced these curves to try to reconcile the complicated movements of the planets, relative to the fixed stars, with a geometry based on the circle. In principle, this is possible! Lagrange (1772) showed that *any* motion along the celestial equator can be approximated arbitrarily closely by epicylic motion, and a more modern version of the result may be found in Sternberg (1969). But Ptolemy's mistake was to accept the apparent complexity of the motions of the planets as actual in the first place. As we now know, the motion becomes simple when one considers motion relative to the sun rather than to the earth and allows orbits to be ellipses.

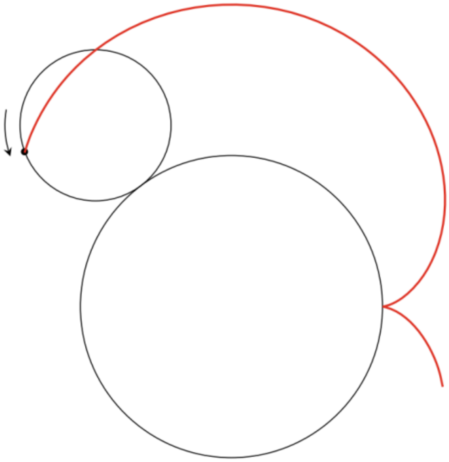

Figure 2.10: Generating an epicycle

Epicycles still have a role to play in engineering, and their mathematical properties are interesting. Some of them are closed curves and turn out to be algebraic, that is, of the form $p(x, y) = 0$ for a polynomial p. Others, such as those that result from rolling circles whose radii have an irrational ratio, lie densely in a certain region of the plane and hence cannot be algebraic; an algebraic curve $p(x, y) = 0$ can meet a straight line $y = mx + c$ in only a finite number of points, corresponding to roots of the polynomial equation $p(x, mx + c) = 0$, and the dense epicycles meet some lines infinitely often.

An obvious relative of the epicycles is the *cycloid*, the curve traced by a point on a circle that rolls on a straight line. The cycloid does not seem to have been studied by the Greeks, but it became a favorite of 17th-century mathematicians. As we will see in Chapter 13, spectacular properties of the cycloid were revealed by the methods of calculus.

Exercises

The equation of the cissoid is derivable as follows.

2.5.1 Using X and Y for the horizontal and vertical coordinates, show that the straight line RP in Figure 2.8 has equation

$$Y = \frac{\sqrt{1 - x^2}}{1 + x}(X - 1).$$

2.5.2 Deduce the equation of the cissoid from Exercise 2.5.1.

The simplest epicyclic curve is the *cardioid* ("heart-shape"), which results from a circle rolling on a fixed circle of the same size.

2.5.3 Sketch a picture of the cardioid, confirming that it is heart-shaped (sort of).

2.5.4 Show that if both circles have radius 1, and we follow the point on the rolling circle initially at $(1, 0)$, then the cardioid it traces out has parametric equations

$$\begin{aligned} x &= 2\cos\theta - \cos 2\theta, \\ y &= 2\sin\theta - \sin 2\theta. \end{aligned}$$

The cardioid is an algebraic curve. Its cartesian equation may be hard to discover, but it is easy to verify, especially if one has a computer algebra system.

2.5.5 Check that the point (x, y) on the cardioid satisfies

$$(x^2 + y^2 - 1)^2 = 4((x - 1)^2 + y^2).$$

3

Greek Number Theory

Preview

Number theory is the second large field of mathematics that comes to us from the Pythagoreans via Euclid. The Pythagorean theorem led mathematicians to the study of squares and sums of squares; Euclid drew attention to the *primes* by proving that there are infinitely many of them.

His investigations were based on the *Euclidean algorithm*, a method for finding the greatest common divisor of two natural numbers. Common divisors are the key to basic results about prime numbers, in particular *unique prime factorization*, which says that each natural number factors into primes in exactly one way.

Another discovery of the Pythagoreans, the irrationality of $\sqrt{2}$, has consequences for natural numbers. Since $\sqrt{2} \neq m/n$ for any natural numbers m, n, there is no integer solution of the equation $x^2 - 2y^2 = 0$. But there are integer solutions of $x^2 - 2y^2 = 1$, and in fact infinitely many of them. The same is true of the equation $x^2 - Ny^2 = 1$ for any nonsquare natural number N.

The latter equation, called *Pell's equation*, is perhaps second in fame only to the Pythagorean equation $x^2 + y^2 = z^2$, among equations for which integer solutions are sought. Equations for which integer or rational solutions are sought are called *Diophantine*, after Diophantus. The methods he used to solve quadratic and cubic Diophantine equations are still of interest. We study his method for cubics in this chapter, and take it up again in Chapter 10.

J. Stillwell, *Mathematics and Its History*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-55193-3_3

The four-square theorem and the pentagonal number theorem were both absorbed around 1830 into Jacobi's theory of theta functions, a much larger theory. Theta functions are related to the *elliptic functions* that we study in Chapter 10.

The *prime numbers* were also considered within the geometric framework, as the numbers with no rectangular representation. A prime number, having no divisors apart from itself and 1, has only a "linear" representation. Of course this is merely a restatement of the definition of prime, and most theorems about prime numbers require much more powerful ideas; however, the Greeks did come up with one gem. This is the proof that there are infinitely many primes, in Book IX of Euclid's *Elements*.

Given any finite collection of primes $p_1, p_2, \ldots, p_n$, we can find another by considering

$$p = p_1 p_2 \cdots p_n + 1.$$

This number is not divisible by $p_1, p_2, \ldots, p_n$ (each leaves remainder 1). Hence either p itself is a prime, and $p > p_1, p_2, \ldots, p_n$, or else it has a prime divisor $\neq p_1, p_2, \ldots, p_n$.

A *perfect number* is one that equals the sum of its divisors (including 1 but excluding itself). For example, $6 = 1 + 2 + 3$ is a perfect number, as is $28 = 1 + 2 + 4 + 7 + 14$. The concept goes back to the Pythagoreans, but only two notable theorems about perfect numbers are known. Euclid concludes Book IX of the *Elements* by proving that if $2^n - 1$ is prime, then $2^{n-1}(2^n - 1)$ is perfect (Exercise 3.2.5). These perfect numbers are of course even, and Euler (1849) (a posthumous publication) proved that every even perfect number is of Euclid's form. Euler's surprisingly simple proof may be found in Burton (1985), p. 504. It is unknown whether odd perfect numbers exist—this may be the oldest open problem in mathematics.

In view of Euler's theorem, all even perfect numbers arise from primes of the form $2^n - 1$. These are known as Mersenne primes, after Marin Mersenne (1588–1648), who first drew attention to the problem of finding primes of this form. It is not known whether there are infinitely many Mersenne primes, though larger and larger ones seem to be found quite regularly. In recent years each new world-record prime has been a Mersenne prime, giving a corresponding world-record perfect number.

Exercises

Infinitely many natural numbers are not sums of three (or fewer) squares. The smallest of them is 7, and it can be shown as follows that no number of the form $8n + 7$ is a sum of three squares.

3.2.1 Show that any square leaves remainder 0, 1, or 4 on division by 8.

3.2.2 Deduce that a sum of three squares leaves remainder 0, 1, 2, 3, 4, 5, or 6 on division by 8.

One reason polygonal numbers play only a small role in mathematics is that questions about them are basically questions about squares—hence the focus is on problems about squares.

3.2.3 Show that the kth pentagonal number is $(3k^2 - k)/2$.

3.2.4 Show that each square is the sum of two consecutive triangular numbers.

Euclid's theorem about perfect numbers depends on the prime divisor property, which will be proved in the next section. Assuming this for the moment, it follows that if $2^n - 1$ is a prime p, then the proper divisors of $2^{n-1}p$ (those unequal to $2^{n-1}p$ itself) are

$$1, 2, 2^2, \ldots, 2^{n-1} \quad \text{and} \quad p, 2p, 2^2 p \ldots, 2^{n-2} p.$$

3.2.5 Given that the divisors of $2^{n-1}p$ are those just listed, show that $2^{n-1}p$ is perfect when $p = 2^n - 1$ is prime.

3.3 The Euclidean Algorithm

This algorithm is named after Euclid because its earliest known appearance is in Book VII of the *Elements*. However, in the opinion of many historians (for example, Heath (1921), p. 399) the algorithm and some of its consequences were probably known earlier. At the very least, Euclid deserves credit for a masterly presentation of the fundamentals of number theory, based on this algorithm.

The Euclidean algorithm is used to find the greatest common divisor (gcd) of two positive integers a, b. The first step is to construct the pair (a_1, b_1), where

$$\begin{aligned} a_1 &= \max(a, b) - \min(a, b), \\ b_1 &= \min(a, b), \end{aligned}$$

and then one simply repeats this operation of subtracting the smaller number from the larger. That is, if the pair constructed at step i is (a_i, b_i), then the pair constructed at step $i + 1$ is

$$\begin{aligned} a_{i+1} &= \max(a_i, b_i) - \min(a_i, b_i), \\ b_{i+1} &= \min(a_i, b_i). \end{aligned}$$

The algorithm terminates at the first stage when $a_{i+1} = b_{i+1}$, and this common value is $\gcd(a, b)$. This is because taking differences preserves any common divisors; hence when $a_{i+1} = b_{i+1}$ we have

$$\gcd(a, b) = \gcd(a_1, b_1) = \cdots = \gcd(a_{i+1}, b_{i+1}) = a_{i+1} = b_{i+1}.$$

The sheer simplicity of the algorithm makes it easy to draw some important consequences. Euclid of course did not use our notation, but nevertheless he had results close to the following.

1. If $\gcd(a, b) = 1$, then there are integers m, n such that $ma + nb = 1$.

   The equations

   $$\begin{aligned} a_1 &= \max(a, b) - \min(a, b), \\ b_1 &= \min(a, b), \\ &\vdots \\ a_{i+1} &= \max(a_i, b_i) - \min(a_i, b_i), \\ b_{i+1} &= \min(a_i, b_i) \end{aligned}$$

   show first that a_1, b_1 are integral linear combinations, $ma + nb$, of a and b, hence so are a_2, b_2, hence so are $a_3, b_3, \ldots$, and finally this is true of $a_{i+1} = b_{i+1}$. But $a_{i+1} = b_{i+1} = 1$, since $\gcd(a, b) = 1$; hence $1 = ma + nb$ for some integers m, n.

2. If p is a prime number that divides ab, then p divides a or b (the *prime divisor property*).

   To see this, suppose p does *not* divide a. Then, since p has no other divisors except 1, we have $\gcd(p, a) = 1$. Hence by the previous result we get integers m, n such that

   $$ma + np = 1.$$

   Multiplying each side by b gives

   $$mab + nbp = b.$$

   By hypothesis, p divides ab; hence p divides *both* terms on the left-hand side, and therefore p divides the right-hand side b.

3. Each positive integer has a unique factorization into primes (the *fundamental theorem of arithmetic*).

   Suppose on the contrary that some integer n has two different prime factorizations:

   $$n = p_1 p_2 \cdots p_j = q_1 q_2 \cdots q_k.$$

   By removing common factors, if necessary, we can assume that there is a p_i that is not among the q's. But this contradicts the previous result, because p_i divides $n = q_1 q_2 \cdots q_k$, yet it does not divide any of $q_1, q_2, \ldots, q_k$ individually, since these are prime numbers $\neq p_i$.

Induction

In this and the previous section we have glossed over an important point that Euclid was aware of but mentioned only briefly—the principle that *an infinite decreasing sequence of positive integers is impossible*. In the present section this *infinite descent* principle guarantees termination of the Euclidean algorithm, necessarily with the number gcd(a, b), for any pair of positive integers a, b. This is because the repeated subtraction process produces steadily decreasing numbers.

In the previous section infinite descent played a hidden role in Euclid's proof that there are infinitely many prime numbers: namely, in the assumption that *some* prime number divides $p_1 p_2 \cdots p_n + 1$. In Proposition 31 of Book VII of his *Elements*, Euclid proves existence of a prime divisor of any number N by repeatedly splitting N into smaller factors. If this process does not arrive at a prime factor then we get an infinite sequence of positive integers, each smaller than the one before. As Euclid says, this is "impossible in numbers."

Today, the impossibility of infinite descent is one way of stating *mathematical induction* (also known as *complete* induction), a method of proof that reflects the nature of positive integers as numbers that arise from 1 by repeatedly adding 1. On the one hand, this property implies that we arrive at 1 from any positive integer by stepping downward only finitely often. On the other hand, it implies that any positive integer can be reached from 1 by finitely often adding 1. In particular, a property P can be proved to hold for all positive integers by proving

1. P holds for the number 1 (the *base step*),

2. If P holds for n, then P holds for $n + 1$ (the *induction step*).

"Base step, induction step" is often considered the standard form of *proof by induction*, but it is perfectly fair to say that proofs by infinite descent, such as Euclid's, are also proofs by induction.

Moreover, it is not generally appreciated that number theory needs induction as much as Euclid needed the parallel axiom in his geometry. The first to appreciate this fact was Grassmann (1861), who showed that all the basic algebraic properties of positive integers, such as $a+b=b+a$ and $ab=ba$, can be proved by induction. Even then, Grassmann's breakthrough was buried in a school textbook, and not brought into general mathematical consciousness until the 1880s, when Peano (1889) formulated an axiom system for arithmetic with an *induction axiom* at its core. This system, called *Peano arithmetic* or PA, is an important part of the foundations of mathematics, as we will see in Chapter 17.

Exercises

We can now fill the gap in the proof of Euclid's theorem on perfect numbers (previous exercise set), using the prime divisor property.

3.3.1 Use the prime divisor property to show that the proper divisors of $2^{n-1}p$, for any odd prime p, are $1, 2, 2^2, \dots, 2^{n-1}$ and $p, 2p, 2^2p \dots, 2^{n-2}p$.

The result that if $\gcd(a,b)=1$ then $1 = ma+nb$ for some integers m and n is a special case of the following way to represent the gcd.

3.3.2 Show that, for any integers a and b, there are integers m and n such that $\gcd(a,b) = ma+nb$.

This in turn gives a general way to find integer solutions of linear equations.

3.3.3 Deduce from Exercise 3.3.2 that the equation $ax+by=c$ with integer coefficients a, b, and c has an integer solution x, y if $\gcd(a,b)$ divides c.

The converse of this result is also valid, as one discovers when considering a *necessary* condition for $ax+by=c$ to have an integer solution.

3.3.4 The equation $12x+15y=1$ has no integer solution. Why?

3.3.5 (Solution of linear Diophantine equations) Give a test to decide, for any given integers a, b, c, whether there are integers x, y such that

$$ax+by=c.$$

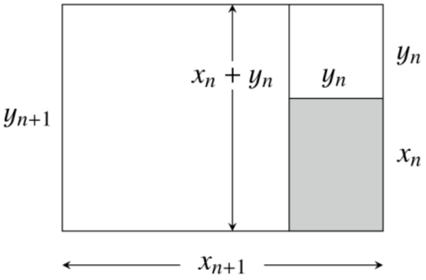

Figure 3.3: The recurrence relation

Nevertheless, one feels that Figure 3.3 gives the most natural interpretation of these relations. The discovery that the same relations generate solutions of $x^2 - 2y^2 = 1$ possibly arose from wishing that the Euclidean algorithm terminated with $x_1 = y_1 = 1$. If the Pythagoreans started with $x_1 = y_1 = 1$ and applied the recurrence relations, then they may well have found that (x_n, y_n) satisfies $x^2 - 2y^2 = (-1)^n$, as we did earlier.

Many other instances of the Pell equation $x^2 - Ny^2 = 1$ occur in Greek mathematics. In the seventh century CE the Indian mathematician Brahmagupta gave a procedure for generating larger solutions of $x^2 - Ny^2 = 1$ from known solutions. But *existence* of a solution, for any non-square N, was rigorously proved only in 1768 by Lagrange. The later European work on Pell's equation, which began in the 17th century with Brouncker and others, was based on the *continued fraction* for $\sqrt{N}$, though this amounts to the same thing as anthyphairesis (see exercises). A short but detailed history of Pell's equation is in Dickson (1920), pp. 341–400.

An interesting aspect of the theory is the very irregular relationship between N and the number of steps before a rectangle proportional to the original recurs. If the number of steps is large, the smallest nontrivial solution of $x^2 - Ny^2 = 1$ is enormous. A famous example is what is called the *cattle problem* of Archimedes (287–212 BCE), which leads to the equation

$$x^2 - 4729494y^2 = 1.$$

Its smallest solution was found by Krummbiegel and Amthor (1880) to have 206,545 digits!

A recent paper on the cattle problem, Lenstra (2002), gives a strikingly condensed form of solution: "for the first time in history, *all* infinitely many solutions to the cattle problem are displayed in a handy little table."

be rational, and its third intersection with the curve will also be rational, by an argument like the preceding one. This fact becomes more useful when one realizes that the two known rational points can be taken to *coincide*, in which case the line is the *tangent* through the known rational point. Thus from one rational solution we can generate another by the tangent construction, and from two we can construct a third by taking the chord between the two.

Diophantus found rational solutions to cubic equations in what seems to have been essentially this way. The surviving works of Diophantus reveal little of his methods, but a plausible reconstruction—an algebraic version of the tangent and chord constructions—has been given by Bashmakova (1981). Probably the first to understand Diophantus's methods was Fermat, in the 17th century, and the first to give the tangent and chord interpretation was Newton (1670s).

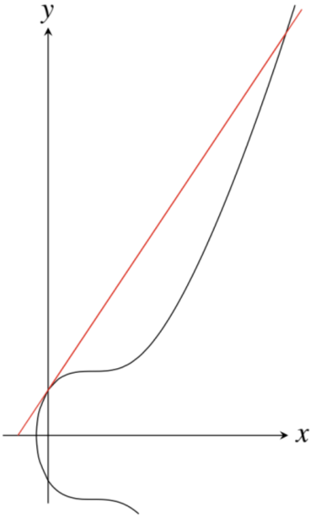

Figure 3.4: Cubic curve $y^2 = x^3 - 3x^2 + 3x + 1$ and tangent

In contrast to the quadratic case, we have no choice in the slope of the rational line for cubics. Thus it is unclear whether this method will give *all* rational points on a cubic. A remarkable theorem, conjectured by Poincaré (1901) and proved by Mordell (1922), says that all rational points can be generated by tangent and chord constructions applied to finitely many points. However, it is still not known whether there is an algorithm for finding a finite set of such rational generators on each cubic curve.

Exercises

3.5.1 Explain the solution $x = 21/4$, $y = 71/8$ to $x^3 - 3x^2 + 3x + 1 = y^2$ given by Diophantus (Heath (1910), p. 242) by constructing the tangent through the obvious rational point on this curve (Figure 3.4).

3.5.2 Rederive the following rational point construction of Viète (1593), p. 145. Given the rational point (a, b) on $x^3 - y^3 = a^3 - b^3$, show that the tangent at (a, b) is

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

and that the other intersection of the tangent with the curve is the rational point

$$x = a\frac{a^3 - 2b^3}{a^3 + b^3}, \qquad y = b\frac{b^3 - 2a^3}{a^3 + b^3}.$$

4

Infinity in Greek Mathematics

Preview

Perhaps the most interesting—and most modern—feature of Greek mathematics is its treatment of infinity. The Greeks feared infinity and tried to avoid it, but in doing so they laid the foundations for a rigorous treatment of infinite processes in 19th century calculus.

The most original contributions to the theory of infinity in ancient times were the *theory of proportions* and the *method of exhaustion*. Both were due to Eudoxus and expounded in Books V and XII of Euclid's *Elements*.

The theory of proportions develops the idea that a "quantity" λ (what we would now call a real number) can be known by its position among the rational numbers. That is, λ is known if we know the rational numbers less than λ and the rational numbers greater than λ. In a sense, the space less than λ can be "exhausted" by rational numbers.

The method of exhaustion generalizes this idea from quantities to regions of the plane or space. A region becomes known (in area or volume) when its position among known areas or volumes is known. For example, we know the area of a circle when we know the areas of the polygons inside it and the areas of polygons outside it; we know the volume of a pyramid when we know the volumes of stacks of prisms inside it and outside it.

Using this method, Euclid found that the volume of a tetrahedron equals 1/3 of its base area times its height, and Archimedes found the area of a parabolic segment. Both of them relied on an infinite process that is fundamental to many calculations of area and volume: the summation of an infinite geometric series.

J. Stillwell, *Mathematics and Its History*, Undergraduate Texts in Mathematics,
https://doi.org/10.1007/978-3-030-55193-3_4