1

Natural numbers and integers

PREVIEW

Counting is presumably the origin of mathematical thought, and it is certainly the origin of difficult mathematical problems. As the great Hungarian problem-solver Paul Erdős liked to point out, if you can think of an open problem that is more than 200 years old, then it is probably a problem in number theory.

In recent decades, difficulties in number theory have actually become a virtue. *Public key encryption*, whose security depends on the difficulty of factoring large numbers, has become one of the commonest applications of mathematics in daily life.

At any rate, problems are the life blood of number theory, and the subject advances by building theories to make them understandable. In the present chapter we introduce some (not so difficult) problems that have played an important role in the development of number theory because they lead to basic methods and concepts.

- Counting leads to *induction*, the key to all facts about numbers, from banalities such as $a+b=b+a$ to the astonishing result of Euclid that there are infinitely many primes.
- Division (with remainder) is the key computational tool in Euclid's proof and elsewhere in the study of primes.
- Binary notation, which also results from division with remainder, leads in turn to a method of "fast exponentiation" used in public key encryption.
- The Pythagorean equation $x^2+y^2=z^2$ from geometry is equally important in number theory because it has integer solutions.

In this chapter we are content to show these ideas at work in few interesting but seemingly random situations. Later chapters will develop the ideas in more depth, showing how they unify and explain a great many astonishing properties of numbers.

1.1 Natural numbers

Number theory starts with the *natural numbers*

$$1,2,3,4,5,6,7,8,9,\ldots,$$

generated from 1 by successively adding 1. We denote the set of natural numbers by $\mathbb{N}$. On $\mathbb{N}$ we have the operations $+$ and $\times$, which are simple in themselves but lead to more sophisticated concepts.

For example, we say that *a divides n* if $n = ab$ for some natural numbers a and b. A natural number p is called *prime* if the only natural numbers dividing p are 1 and p itself.

Divisibility and primes are behind many of the interesting questions in mathematics, and also behind the recent applications of number theory (in cryptography, internet security, electronic money transfers etc.).

The sequence of prime numbers begins with

$$2,3,5,7,11,13,17,19,23,29,31,37,\ldots$$

and continues in a seemingly random manner. There is so little pattern in the sequence that one cannot even see clearly whether it continues forever. However, Euclid (around 300 BCE) proved that *there are infinitely many primes*, essentially as follows.

Infinitude of primes. *Given any primes $p_1, p_2, p_3, \ldots, p_k$, we can always find another prime p.*

Proof. Form the number

$$N = p_1 p_2 p_3 \cdots p_k + 1.$$

Then none of the given primes $p_1, p_2, p_3, \ldots, p_k$ divides N because they all leave remainder 1. On the other hand, *some* prime p divides N. If N itself is prime we can take $p = N$, otherwise $N = ab$ for some smaller numbers a and b. Likewise, if either a or b is prime we take it to be p, otherwise split a and b into smaller factors, and so on. Eventually we must reach a prime p dividing N because natural numbers cannot decrease forever. □

Exercises

Not only is the sequence of primes without apparent pattern, there is not even a known simple formula that produces only primes. There are, however, some interesting "near misses".

1.1.1 Check that the quadratic function n^2+n+41 is prime for all small values of n (say, for n up to 30).

1.1.2 Show nevertheless that n^2+n+41 is not prime for certain values of n.

1.1.3 Which is the smallest such value?

1.2 Induction

The method just used to find the prime divisors of N is sometimes called *descent*, and it is an instance of a general method called *induction*.

The "descent" style of induction argument relies on the fact that any process producing smaller and smaller natural numbers must eventually halt. The process of repeatedly adding 1 reaches any natural number n in a finite number of steps, hence there are only finitely many steps *downward* from n. There is also an "ascent" style of induction that imitates the construction of the natural numbers themselves—starting at some number and repeatedly adding 1.

An "ascent" induction proof is carried out in two steps: the *base step* (getting started) and the *induction step* (going from n to $n+1$). Here is an example: proving that *any number of the form* k^3+2k *is divisible by* 3.

Base step. The claim is true for $k=1$ because $1^3+2\times 1=3$, which is certainly divisible by 3.

Induction step. Suppose that the claim is true for $k=n$, that is, 3 divides n^3+2n. We want to deduce that it is true for $k=n+1$, that is, that 3 divides $(n+1)^3+2(n+1)$. Well,

$$\begin{aligned}&(n+1)^3+2(n+1)\\ &=n^3+3n^2+3n+1+2n+2\\ &=n^3+2n+3n^2+3n+3\\ &=n^3+2n+3(n^2+n+1)\end{aligned}$$

And the right-hand side is the sum of n^3+2n, which we are supposing to be divisible by 3, and $3(n^2+n+1)$, which is obviously divisible by 3. Therefore $(n+1)^3+2(n+1)$ is divisible by 3, as required. □

Induction is fundamental not only for proofs of theorems about $\mathbb{N}$ but also for defining the basic functions on $\mathbb{N}$. Only one function needs to be assumed, namely the *successor function* $s(n) = n+1$; then $+$ and $\times$ can be defined by induction. In this book we are not trying to build everything up from bedrock, so we shall assume $+$ and $\times$ and their basic properties, but it is worth mentioning their inductive definitions, since they are so simple.

For any natural number m we define $m+1$ by

$$m+1 = s(m).$$

Then, given the definition of $m+n$ for all m, we define $m+s(n)$ by

$$m+s(n) = s(m+n).$$

It then follows, by induction on n, that $m+n$ is defined for all natural numbers m and n. The definition of $m \times n$ is similarly based on the successor function and the $+$ function just defined:

$$m \times 1 = m$$
$$m \times s(n) = m \times n + m.$$

From these inductive definitions one can give inductive *proofs* of the basic properties of $+$ and $\times$, for example $m+n = n+m$ and $l(m+n) = lm+ln$. Such proofs were first given by Grassmann (1861) (in a book intended for high school students!) but they went unnoticed. They were rediscovered, together with an analysis of the successor function itself, by Dedekind (1888). For more on this see Stillwell (1998), Chapter 1.

Exercises

An interesting process of descent may be seen in the algorithm for the so-called *Egyptian fractions* introduced by Fibonacci (1202). The goal of the algorithm is to represent any fraction $\frac{b}{a}$ with $0 < b < a$ as sum of distinct terms $\frac{1}{n}$, called *unit fractions*. (The ancient Egyptians represented fractions in this way.)

Fibonacci's algorithm, in a nutshell, is to *repeatedly subtract the largest possible unit fraction*. Applied to the fraction $\frac{11}{12}$, for example, it yields

$$\frac{11}{12} - \frac{1}{2} = \frac{5}{12}, \quad \text{subtracting the largest unit fraction, } \tfrac{1}{2}, \text{ less than } \tfrac{11}{12},$$
$$\frac{5}{12} - \frac{1}{3} = \frac{1}{12}, \quad \text{subtracting the largest unit fraction, } \tfrac{1}{3}, \text{ less than } \tfrac{5}{12},$$
$$\text{hence } \frac{11}{12} = \frac{1}{2} + \frac{1}{3} + \frac{1}{12}.$$

It turns out that the fractions produced by the successive subtractions always have a descending sequence of numerators (11, 5, 1 in the example), hence they necessarily terminate with 1.

1.2.1 Use Fibonacci's algorithm to find an Egyptian representation of $\frac{9}{11}$.

1.2.2 If a, b, q are natural numbers with $\frac{1}{q+1} < \frac{b}{a} < \frac{1}{q}$, show that

$$\frac{b}{a} - \frac{1}{q+1} = \frac{b'}{a(q+1)} \quad \text{where} \quad 0 < b' < b.$$

Hence explain why Fibonacci's algorithm always works.

1.3 Integers

For several reasons, it is convenient to extend the set $\mathbb{N}$ of natural numbers to the *group* $\mathbb{Z}$ of *integers* by throwing in the *identity* element 0 and an *inverse* $-n$ for each natural number n. One reason for doing this is to ensure that the difference $m-n$ of any two integers is meaningful. Thus $\mathbb{Z}$ is a set on which all three operations $+$, $-$, and $\times$ are defined. (The notation $\mathbb{Z}$ comes from the German "Zahlen", meaning "numbers".)

$\mathbb{Z}$ is an *abelian group* under the operation $+$, because it has the three group properties:

Associativity: $a+(b+c) = (a+b)+c$
Identity: $a+0 = a$
Inverse: $a+(-a) = 0$

and also the abelian property: $a+b = b+a$.

$\mathbb{Z}$ is much older than the concept of abelian group. The latter concept could only be conceived after other examples came to light, particularly *finite* abelian groups. We shall meet some of them in Chapter 3.

$\mathbb{Z}$ is a *ring* under the operations $+$ and $\times$: it is an abelian group under $+$ and the $\times$ is linked with $+$ by

Distributivity: $a(b+c) = ab+ac$.

The ring concept also emerged much later than $\mathbb{Z}$. It grew out of 18th and 19th century attempts to generalize the concept of integer. We see one of these in Section 1.8, and take up the general ring concept in Chapter 10.

The ring properties show that $\mathbb{Z}$ has more structure than $\mathbb{N}$, though it must be admitted that this does not make everything simpler. The presence

of the negative integers $-1,-2,-3,\ldots$ in $\mathbb{Z}$ slightly complicates the concept of prime number. Since any integer n is divisible by 1, -1, n and $-n$, we have to define a *prime* in $\mathbb{Z}$ to be an integer p divisible only by ± 1 (the so-called *units* of $\mathbb{Z}$) and $\pm p$.

In general, however, it is simpler to work with integers than natural numbers. Here is a problem that illustrates the difference.

Problem. Describe the numbers $4m+7n$

1. where m and n are natural numbers,
2. where m and n are integers.

In the first case the numbers are 11, 15, 18, 19, 22, 23, 25, 26, 27 and all numbers ≥ 29. The numbers < 29 can be verified (laboriously, I admit) by trial. To see why all numbers ≥ 29 are of the form $4m+7m$, we first verify this for 29, 30, 31, 32; namely

$$\begin{aligned} 29 &= 2\times 4+3\times 7 \\ 30 &= 4\times 4+2\times 7 \\ 31 &= 6\times 4+1\times 7 \\ 32 &= 1\times 4+4\times 7. \end{aligned}$$

Then we can get the next four natural numbers by adding one more 4 to each of these, then the next four by adding two more 4s, and so on (this is really an induction proof).

In the second case, all integers are obtainable. This is simply because $1 = 4\times 2-7$, and therefore $n = 4\times 2n-7\times n$, for any integer n.

This type of problem is easier to understand with the help of the gcd—*greatest common divisor*—which we study in the next chapter. But first we need to look more closely at division, particularly division with remainder, which is the subject of the next section.

Exercises

A concrete problem similar to describing $4m+7n$ is the *McN*ggets problem*: given that McN*ggets can be bought in quantities of 6, 9 or 20, which numbers of McN*ggets can be bought? This is the problem of describing the numbers $6i+9j+20k$ for natural numbers or zero i, j and k.

It turns out the possible numbers include all numbers ≥ 44, and an irregular set of numbers < 43.

1.3.1 Explain why the number 43 is not obtainable.

1.3.2 Show how each of the numbers 44, 45, 46, 47, 48, 49 is obtainable.

1.3.3 Deduce from Exercise 1.3.2 that any number > 43 is obtainable.

But if the negative quantities -6, -9 and -20 are allowed (say, by selling McN*ggets back), then any integer number of McN*ggets can be obtained.

1.3.4 Show in fact that $1 = 9m + 20n$ for some integers m and n.

1.3.5 Deduce from Exercise 1.3.4 that every integer is expressible in the form $9m + 20n$, for some integers m and n.

1.3.6 Is every integer expressible in the form $6m + 9n$? What do the results in Exercises 1.3.4 and 1.3.5 have to do with common divisors?

1.4 Division with remainder

As mentioned in Section 1.1, a natural number b is said to *divide* n if $n = bc$ for some natural number c. We also say that b is a *divisor of* n, and that n is a *multiple of* b. The same definitions apply wherever there is a concept of multiplication, such as in $\mathbb{Z}$.

In $\mathbb{N}$ or $\mathbb{Z}$ it may very well happen that b does *not* divide a, for example, 4 does not divide 23. In this case we are interested in the *quotient* q and *remainder* r when we do *division of* a *by* b. The quotient comes from the greatest multiple qb of b that is $\leq a$, and the remainder is $a - qb$. For example

$$23 = 5 \times 4 + 3,$$

so when we divide 23 by 4 we get quotient 5 and remainder 3.

The remainder $r = a - qb$ may be found by repeatedly subtracting b from a. This gives natural numbers $a, a-b, a-2b, \ldots$, which decrease and therefore include a least member $r = a - qb \geq 0$ by descent. Then $r < b$, otherwise we could subtract b again. The remainder $r < b$ is also evident in Figure 1.1, which shows a lying between successive multiples of b, hence necessarily at distance $< b$ from the nearest such multiple, qb.

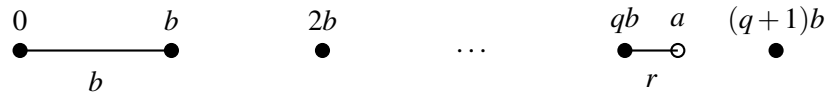

Figure 1.1: Division with remainder

Important. The main purpose of division with remainder is to find the remainder, which tells us whether b divides a or not.

It does not help (and it may be confusing) to form the fraction a/b, because this brings us no closer to knowing whether b divides a. For example, the fraction

$$\frac{43560029}{77777}$$

does not tell us whether 77777 divides 43560039 or not. To find out, we need to know whether the remainder is 0 or not. We could do the full division with remainder:

$$43560029 = 560 \times 77777 + 4909$$

which tells us the exact remainder, 4909, or else evaluate the fraction numerically

$$\frac{43560029}{77777} = 560.0631...,$$

which is enough to tell us that the remainder is $\neq 0$. (And we can read off the quotient $q = 560$ as the part before the decimal point, and hence find the remainder, as $43560029 - 560 \times 77777 = 4909$.)

Exercises

1.4.1 Using a calculator or computer, use the method above to find the remainder when 12345678 is divided by 3333.

1.4.2 Calculate the multiples of 3333 on either side of 12345678.

1.5 Binary notation

Division with remainder is the natural way to find the *binary numeral* of any natural number n. The digits of the numeral are found by dividing n by 2, writing the remainder, and repeating the process with the quotient until the quotient 0 is obtained. Then the sequence of remainders, written in reverse order, is the binary numeral for n.

Example. Binary numeral for 2001.

$$\begin{aligned}
2001 &= 1000 \times 2 + 1 \\
1000 &= 500 \times 2 + 0 \\
500 &= 250 \times 2 + 0 \\
250 &= 125 \times 2 + 0 \\
125 &= 62 \times 2 + 1 \\
62 &= 31 \times 2 + 0 \\
31 &= 15 \times 2 + 1 \\
15 &= 7 \times 2 + 1 \\
7 &= 3 \times 2 + 1 \\
3 &= 1 \times 2 + 1 \\
1 &= 0 \times 2 + 1.
\end{aligned}$$

Hence the binary numeral for 2001 is 11111010001.

A general binary numeral $a_k a_{k-1} \ldots a_1 a_0$, where each a_i is 0 or 1, stands for the number

$$n = a_k 2^k + a_{k-1} 2^{k-1} + \cdots + a_1 2 + a_0,$$

because repeated division of this number by 2 yields the successive remainders $a_0, a_1, \ldots, a_{k-1}, a_k$. Thus one can reconstruct n from its binary digits by multiplying them by the appropriate powers of 2 and adding.

However, it is more efficient to view $a_k a_{k-1} \ldots a_1 a_0$ as a code for constructing n from the number 0 by a sequence of doublings (multiplications by 2) and additions of 1, namely the *reverse of the sequence of operations by which the binary numeral was computed from n*. Moving from left to right, one doubles and adds a_i (if nonzero) for each digit a_i.

Figure 1.2 shows a way to set out the computation, recovering 2001 from its binary numeral 11111010001.

The number of operations

The number of doublings in this process is one less than the number of digits in the binary numeral for n, hence less than $\log_2 n$, since the largest number with k digits is $2^k - 1$ (whose binary numeral consists of k ones), and its log to base 2 is therefore $< k = \log_2(2^k)$.

1	1	1	1	1	0	1	0	0	0	1		
+1											=	1
× 2											=	2
	+1										=	3
	×2										=	6
		+1									=	7
		×2									=	14
			+1								=	15
			×2								=	30
				+1							=	31
				×2							=	62
					+0						=	62
					×2						=	124
						+1					=	125
						×2					=	250
							+0				=	250
							×2				=	500
								+0			=	500
								×2			=	1000
									+0		=	1000
									×2		=	2000
										+1	=	2001

Figure 1.2: Recovering a number from its binary numeral

Likewise, there are $< \log_2 n$ additions. So the total number of operations, either doubling or adding 1, needed to produce n is $< 2\log_2 n$.

This observation gives a highly efficient way to compute powers, based on *repeated squaring*. To form m^n, we begin with $m = m^1$, and repeatedly double the exponent (by squaring) or add 1 to it (by multiplying by m). Since we can reach exponent n by doubling or adding 1 less than $2\log_2 n$ times, we can form m^n by squaring or multiplying by m less than $2\log_2 n$ times. That is, *it takes less than* $2\log_2 n$ *multiplications to form* m^n.

Thus the number of operations is roughly proportional to the length of n (the number of its binary or decimal digits). Few problems in number theory can be solved in so few steps, and the fast solution of this particular problem is crucial in modern cryptography and electronic security systems (see Chapter 4).

Exercises

Binary notation is more often used by computers than humans, since we have 10 fingers and hence find it convenient to use base 10 rather than base 2. However, some famous numbers are most simply written in binary. Examples are the *Mersenne primes*, which are prime numbers of the form $2^p - 1$ where p is prime.

1.5.1 Show that the binary numeral for $2^p - 1$ is $111\cdots1$ (p digits), and that the first four Mersenne primes have binary numerals 11, 111, 11111, and 1111111.

1.5.2 However, not every prime p gives a prime $2^p - 1$: factorize $2^{11} - 1$.

1.5.3 Show also that $2^n - 1$ is *never* a prime when n is not prime. (*Hint*: suppose that $n = pq$, let $x = 2^p$, and show that $x - 1$ divides $x^q - 1$.)

Mersenne primes are named after Marin Mersenne (1588–1648) who first drew attention to the problem of finding them. They occur (though not under that name) in a famous theorem of Euclid on *perfect numbers*. A number is called perfect if it equals the sum of its proper divisors (divisors less than itself). For example, 6 is perfect, because its proper divisors are 1, 2 and 3, and $6 = 1 + 2 + 3$. Euclid's theorem is: *if* $2^p - 1$ *is prime then* $2^{p-1}(2^p - 1)$ *is perfect.*

We discuss this theorem further in Chapter 2 when we have developed some theory of divisibility. In the meantime we observe that Euclid's perfect numbers also have binary numerals of a simple form.

1.5.4 Show that the first four perfect numbers arising from Mersenne primes have binary numerals 110, 11100, 111110000, and 1111111000000.

1.5.5 What is the binary numeral for $2^{p-1}(2^p - 1)$?

1.6 Diophantine equations

Solving equations is the traditional goal of algebra, and particular parts of algebra have been developed to analyze particular methods of solution. *Solution by radicals* is one branch of the tradition, typified by the ancient formula

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

for the solution of the general quadratic equation $ax^2 + bx + c = 0$, and by more complicated formulas (involving cube roots as well as square roots) for the solution of cubic and quartic equations. This method of solution is analyzed by means of the *field* and *group* concepts, which lead to *Galois theory*. Its main results may be found in the companion book to this one, Stillwell (1994).

The other important branch of the tradition is *finding integer solutions*, the main theme of the present book. It leads to the *ring* concept and *ideal theory*. Equations whose integer solutions are sought are called *Diophantine*, even though it is not really the equations that are "Diophantine", but the solutions. Nevertheless, certain equations stand out as "Diophantine" because their integer solutions are of exceptional interest.

- The Pythagorean equation $x^2 + y^2 = z^2$, whose natural number solutions (x, y, z) are known as *Pythagorean triples.*
- The Pell equation $x^2 - ny^2 = 1$ for any nonsquare natural number n.
- The Bachet equation $y^3 = x^2 + n$ for any natural number n.
- The Fermat equation $x^n + y^n = z^n$ for any integer $n > 2$.

The Pythagorean equation is the oldest known mathematical problem, being the subject of a Babylonian clay tablet from around 1800 BCE known as Plimpton 322 (from its museum catalogue number). The tablet contains the two columns of natural numbers, y and z shown in Figure 1.3.

y	z
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: Plimpton 322

The left part of the table is missing, but it is surely a column of values of x, because each value of $z^2 - y^2$ is an integer square x^2, and so the table is essentially a list of Pythagorean triples.

This means that Pythagorean triples were known long before Pythagoras (who lived around 500 BCE), and the Babylonians apparently had sophisticated means of producing them. Notice that Plimpton 322 does not contain any well known Pythagorean triples, such as $(3,4,5)$, $(5,12,13)$ or $(8,15,17)$. It does, however, contain triples derived from these, mostly in nontrivial ways.

Around 300 BCE, Euclid showed that all natural number solutions of $x^2 + y^2 = z^2$ can be produced by the formulas

$$x = (u^2 - v^2)w, \quad y = 2uvw, \quad z = (u^2 + v^2)w$$

by letting u, v and w run through all the natural numbers. (Also the same formulas with x and y interchanged.)

It is easily checked that these formulas give

$$x^2 + y^2 = z^2,$$

but it is not so easily seen that every solution is of Euclid's form. Another approach, using rational numbers, was found by Diophantus around 200 CE. Diophantus specialized in solving equations in rationals, so his solutions are not properly "Diophantine" in our sense, but in this case rational and integer solutions are essentially equivalent.

Exercises

1.6.1 Check (preferably with the help of computer) that $z^2 - y^2$ is a perfect square for each pair (y,z) in Plimpton 322.

1.6.2 Check also that x is a "round" number in the Babylonian sense, that is generally divisible by 60, or at least by a divisor of 60. (The Babylonian system of numerals had base 60.)

1.6.3 Verify that if

$$x = (u^2 - v^2)w, \quad y = 2uvw, \quad z = (u^2 + v^2)w$$

then $x^2 + y^2 = z^2$.

1.6.4 Find values of u and v (with $w = 1$) that yield the Pythagorean triples (3,4,5), (5,12,13), (7,24,25) and (8,15,17) when substituted in Euclid's formulas.

1.7 The Diophantus chord method

An integer solution $(x,y,z) = (a,b,c)$ of $x^2+y^2=z^2$ implies

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

so $X = a/c$, $Y = b/c$ is a *rational* solution of the equation

$$X^2+Y^2 = 1,$$

in other words, a *rational point* on the unit circle. (Admittedly, any multiple of the triple, (ma,mb,mc), corresponds to the same point, but we can easily insert multiples once we have found a, b and c from X and Y.)

Diophantus found rational points on $X^2+Y^2=1$ by an algebraic method, which has the geometric interpretation shown in Figure 1.4.

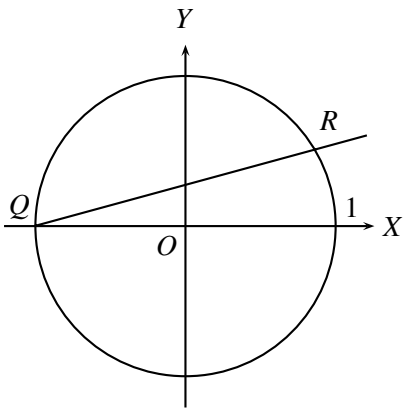

Figure 1.4: The chord method for rational points

If we draw the chord connecting an arbitrary rational point R to the point $Q = (-1,0)$ we get a line with rational slope, because the coordinates of R and Q are rational. If the slope is t, the equation of this line is

$$Y = t(X+1).$$

Conversely, any line of this form, with rational slope t, meets the circle at a rational point R. This can be seen by computing the coordinates of R. We do this by substituting $Y = t(X+1)$ in $X^2+Y^2=1$, obtaining

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

which is the following quadratic equation for X:

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

The quadratic formula gives the solutions

$$X=-1,\frac{1-t^2}{1+t^2}.$$

The solution $X=-1$ corresponds to the point Q, so the X coordinate at R is $\frac{1-t^2}{1+t^2}$, and hence the Y coordinate is

$$Y=t\left(\frac{1-t^2}{1+t^2}+1\right)=\frac{2t}{1+t^2}.$$

To sum up: an arbitrary rational point on the unit circle $X^2+Y^2=1$ has coordinates

$$\left(\frac{1-t^2}{1+t^2},\frac{2t}{1+t^2}\right),\quad \text{for arbitrary rational } t.$$

Now we can recover Euclid's formulas.

An arbitrary rational t can be written $t=v/u$ where $u,v\in\mathbb{Z}$, and the rational point R then becomes

$$\left(\frac{1-\frac{v^2}{u^2}}{1+\frac{v^2}{u^2}},\frac{2\frac{v}{u}}{1+\frac{v^2}{u^2}}\right)=\left(\frac{u^2-v^2}{u^2+v^2},\frac{2uv}{u^2+v^2}\right).$$

Thus if this is

$$\left(\frac{x}{z},\frac{y}{z}\right)\quad \text{for some } x,y,z\in\mathbb{Z}$$

we must have

$$\frac{x}{z}=\frac{u^2-v^2}{u^2+v^2},\quad \frac{y}{z}=\frac{2uv}{u^2+v^2}$$

for some $u,v\in\mathbb{Z}$.

Euclid's formulas for x, y and z also give these formulas for x/z and y/z, so the results of Euclid and Diophantus are essentially the same.

There is little difference between rational and integer solutions of the equation $x^2+y^2=z^2$ because it is *homogeneous* in x, y and z, hence any rational solution can be multiplied through to give an integer solution. The

situation is quite different with inhomogeneous equations, such as $y^2 = x^3 - 2$, where the integer solutions may be much harder to find.

Diophantus' method for rational solutions can be generalized to cubic equations, where it has enjoyed great success. See for example, Silverman and Tate (1992). However, it does not yield integer solutions except in the rare cases where the equation is homogeneous, and hence it diverges from the path we follow in this book. Indeed, it is often the case—for example with Bachet equations—that a cubic equation has infinitely many rational solutions and only finitely many integer solutions. Since we wish to study integer solutions, we now take our leave of the chord construction, and turn in the next section to an algebraic approach to Pythagorean triples: the use of "generalized integers".

Exercises

Diophantus himself extended his method to equations of the form

$$y^2 = \text{cubic in } x,$$

where all coefficients are rational. Here the link between the geometry and the algebra is that *a straight line through two rational points meets the curve in a third rational point*. When there is only one "obvious" rational point on the curve, then one can use the tangent through this point instead of a chord, because the tangent meets the curve twice when viewed algebraically.

The equation $y^2 = x^3 - 2$ is a good one to illustrate the tangent method, as well as the formidable calculations it can lead to. (Note that this is a Bachet equation; here we have interchanged x and y to conform with the usual notation for cubic curves.)

1.7.1 Show that the tangent to $y^2 = x^3 - 2$ at the "obvious" rational point $(3,5)$ is $y = \frac{27x}{10} - \frac{31}{10}$.

1.7.2 By substituting $y = \frac{27x}{10} - \frac{31}{10}$ in the equation of the curve, show that the tangent meets the curve where $100x^3 - 729x^2 + 1674x - 1161 = 0$.

1.7.3 By dividing $100x^3 - 729x^2 + 1674x - 1161$ twice by $x - 3$, or otherwise, show that the tangent meets the curve twice at $x = 3$ and once at $x = \frac{129}{100}$.

1.7.4 Hence find a rational point on $y^2 = x^3 - 2$ other than $(3, \pm 5)$.

There are in fact infinitely many rational points on the curve $y^2 = x^3 - 2$ (though this was not known until 1930; see Mordell (1969), Chapter 26), but we show later that its only integer points are $(3, \pm 5)$.

1.8 Gaussian integers

The Pythagorean equation appears in a new light if we use complex numbers to factorize the sum of two squares:

$$x^2+y^2=(x-yi)(x+yi) \quad \text{where } i=\sqrt{-1}.$$

Given that x and y are integers, the factors $x-yi$, $x+yi$ may be regarded as "complex integers". We denote the set of such "integers" by

$$\mathbb{Z}[i]=\{a+bi : a,b\in\mathbb{Z}\}$$

and call them the *Gaussian integers*, after Gauss, who was the first to realize that $\mathbb{Z}[i]$ has many properties in common with $\mathbb{Z}$.

For a start, it is clear that the sum, difference and product of numbers in $\mathbb{Z}[i]$ are also in $\mathbb{Z}[i]$, hence we can freely use $+$, $-$, and $\times$ and calculate by the same rules as in $\mathbb{Z}$. This already gives nice results about sums of squares and Pythagorean triples.

Two square identity. *A sum of two squares times a sum of two squares is a sum of two squares, namely*

$$(a_1^2+b_1^2)(a_2^2+b_2^2)=(a_1a_2-b_1b_2)^2+(a_1b_2+b_1a_2)^2.$$

Proof. We factorize the sums of two squares as above, then recombine the two factors with negative signs, and the two factors with positive signs:

$$\begin{aligned}(a_1^2+b_1^2)(a_2^2+b_2^2)&=(a_1-b_1i)(a_1+b_1i)(a_2-b_2i)(a_2+b_2i)\\&=(a_1-b_1i)(a_2-b_2i)(a_1+b_1i)(a_2+b_2i)\\&=[a_1a_2-b_1b_2-(a_1b_2+b_1a_2)i]\times\\&\qquad\qquad[a_1a_2-b_1b_2+(a_1b_2+b_1a_2)i]\\&=(a_1a_2-b_1b_2)^2+(a_1b_2+b_1a_2)^2.\end{aligned}$$

□

Corollary. *If the triples (a_1,b_1,c_1) and (a_2,b_2,c_2) are Pythagorean, then so is the triple $(a_1a_2-b_1b_2,a_1b_2+b_1a_2,c_1c_2)$.*

Proof. If (a_1,b_1,c_1) and (a_2,b_2,c_2) are Pythagorean triples, then

$$a_1^2+b_1^2=c_1^2 \quad \text{and} \quad a_2^2+b_2^2=c_2^2.$$

It follows that

$$\begin{aligned}(c_1c_2)^2 = c_1^2c_2^2 &= (a_1^2+b_1^2)(a_2^2+b_2^2)\\ &= (a_1a_2-b_1b_2)^2+(a_1b_2+b_1a_2)^2 \quad \text{by the identity above,}\end{aligned}$$

and this says that $(a_1a_2-b_1b_2, a_1b_2+b_1a_2, c_1c_2)$ is a Pythagorean triple. □

Of course, the two square identity can be proved without using $\sqrt{-1}$, by multiplying out both sides and comparing the results. And presumably it was first discovered this way, because it was known long before the introduction of complex numbers. Though first given explicitly by al-Khazin around 950 CE, it seems to have been known to Diophantus, and perhaps even to the Babylonians, because many of the triples implicit in Plimpton 322 can be obtained from smaller triples by the Corollary (see exercises).

However, the two square identity is more natural in the world $\mathbb{C}$ of complex numbers because it expresses one of their fundamental properties: namely, the *multiplicative property of their norm*. If $z = a+bi$ we define

$$\text{norm}(z) = |a+bi|^2 = a^2+b^2,$$

and it follows from the two square identity that

$$\text{norm}(z_1)\text{norm}(z_2) = \text{norm}(z_1z_2) \qquad (*)$$

because $z_1 = a_1+b_1i$ and $z_2 = a_2+b_2i$ imply

$$z_1z_2 = a_1a_2-b_1b_2+(a_1b_2+b_1a_2)i.$$

In algebra and complex analysis it is more common to state the multiplicative property (*) in terms of the *absolute value* $|z| = \sqrt{a^2+b^2}$, namely

$$|z_1||z_2| = |z_1z_2|. \qquad (**)$$

(*) and (**) are obviously equivalent, but the norm is the more useful concept in $\mathbb{Z}[i]$ because it is an ordinary integer, and this allows certain properties of $\mathbb{Z}[i]$ to be derived from properties of $\mathbb{Z}$.

So much for the elementary properties of Gaussian integers. $\mathbb{Z}[i]$ also has deeper properties in common with $\mathbb{Z}$, involving divisors and primes. These properties will be proved for $\mathbb{Z}$ in the next chapter, and for $\mathbb{Z}[i]$ in Chapter 6.

However, we can travel a little further in the right direction by following the dream that $\mathbb{Z}[i]$ holds the secrets of the Pythagorean equation

$$z^2 = x^2 + y^2 = (x - yi)(x + yi).$$

If the integers x and y have no common prime divisor, then it seems likely that $x - yi$ and $x + yi$ also have no common prime divisor, whatever "prime" means in $\mathbb{Z}[i]$. If so, then it would seem that the factors $x - yi$, $x + yi$ of the square z^2 are *themselves squares* in $\mathbb{Z}[i]$. In particular,

$$x - yi = (u - vi)^2 \quad \text{for some } u, v \in \mathbb{Z}.$$

But in that case

$$x - yi = (u^2 - v^2) - 2uvi$$

and, equating real and imaginary parts,

$$x = u^2 - v^2, \quad y = 2uv, \quad \text{and hence} \quad z = u^2 + v^2.$$

Thus we have arrived again at Euclid's formula for Pythagorean triples! (Or more precisely, the formula for *primitive* Pythagorean triples, from which all others are obtained as constant multiples. The primitive triples are those for which x, y, and z have no common prime divisor, and they result from u and v with no common prime divisor.)

The idea that factors of a square with no common prime divisor are themselves squares is essentially correct in $\mathbb{Z}[i]$, but to see why we must first understand why it is correct in $\mathbb{N}$. This will be explained in the next chapter.

Exercises

The rule in the Corollary for generating new Pythagorean triples from old gives some interesting results.

1.8.1 Find the Pythagorean triples generated from

- (4,3,5) and itself,
- (12,5,13) and itself,
- (15,8,17) and itself.

1.8.2 Do these results account for any of the entries in Plimpton 322?

1.8.3 Try to generate other entries in Plimpton 322 from smaller triples.

It is clear that we can generate infinitely many Pythagorean triples (x, y, z) but not clear (even from Euclid's formulas) whether there are any significant constraints on their members x, y, and z. For example, can we have x and y odd and z even? This question can be answered by considering remainders on division by 4.

1.8.4 Show that the square of an odd integer $2n+1$ leaves remainder 1 on division by 4.

1.8.5 What is the remainder when an even square is divided by 4?

1.8.6 Deduce from Exercises 1.8.4 and 1.8.5 that the sum of odd squares is never a square.

1.9 Discussion

The discovery of Pythagorean triples, in which the sum x^2+y^2 of two squares is itself a square, leads to a more general question: what values are taken by x^2+y^2 as x and y run through $\mathbb{Z}$? The exercises above imply that x^2+y^2 can *not* take a value of the form $4n+3$ (why?), and the main problem in describing its possible values is to find the *primes* of the form x^2+y^2.

Such questions were first studied by Fermat around 1640, sparked by his reading of Diophantus. He was able to answer them, and also the corresponding questions for x^2+2y^2 and x^2+3y^2. In the 18th century this led to study of the general *quadratic form* $ax^2+bxy+cy^2$ by Euler, Lagrange, Legendre and Gauss. The endpoint of these investigations was the *Disquisitiones Arithmeticae* of Gauss (1801), a book of such depth and complexity that the best number theorists of the 19th century—Dirichlet, Kummer, Kronecker, and Dedekind—found that they had to rewrite it so that ordinary mortals could understand Gauss's results.

The reason that the *Disquisitiones* is so complex is that abstract algebra did not exist when Gauss wrote it. Without new algebraic concepts the deep structural properties of quadratic forms discovered by Gauss cannot be clearly expressed; they can barely be glimpsed by readers lacking the technical power of Gauss. It was precisely to comprehend Gauss's ideas and convey them to others that Kummer, Kronecker, and Dedekind introduced the concepts of rings, ideals, and abelian groups.

An intermediate step in the evolution of ring theory was the creation of *algebraic number theory*: a theory in which algebraic numbers such as $\sqrt{2}$ and i are used to illuminate the properties of natural numbers and integers. Around 1770, Euler and Lagrange had already used algebraic

numbers to study certain Diophantine equations. For example, Euler successfully found all the integer solutions of $y^3 = x^2 + 2$ by factorizing the right-hand side into $(x + \sqrt{-2})(x - \sqrt{-2})$. He assumed that numbers of the form $a + b\sqrt{-2}$ "behave like" integers when a and b themselves are integers (see Section 7.1). The same assumption enables one to determine all primes of the form $x^2 + 2y^2$.

Such reasoning was rejected by Gauss in the *Disquisitiones*, since it was not sufficiently clear what it meant for algebraic numbers to "behave like" integers. In 1801 Gauss may already have known systems of algebraic numbers that did *not* behave like the integers. He therefore worked directly with quadratic forms and their integer coefficients, subduing them with his awesome skill in traditional algebra. However, Gauss (1832) took the first step towards an abstract theory of *algebraic integers* by proving that the Gaussian integers $\mathbb{Z}[i]$ do indeed "behave like" the ordinary integers $\mathbb{Z}$, specifically with respect to prime factorization. Among other things, this gives an elegant way to treat the quadratic form $x^2 + y^2$, as we see in Chapter 6.

The great achievement of Kummer and Dedekind was to tame the systems of algebraic numbers that do *not* behave like $\mathbb{Z}$, by adjoining new "numbers" to them. Kummer's mystery *ideal numbers*, and Dedekind's demystification of them in 1871, are among the most dramatic discoveries of mathematics. Ideal numbers also emerge naturally from the theory of quadratic forms, in particular from the form $x^2 + 5y^2$, so we follow the thread of quadratic forms throughout this book. Quadratic forms not only give the correct historical context for most of the concepts normally covered in ring theory but also provide the simplest and clearest examples.

Springer

http://www.springer.com/978-0-387-95587-2

Elements of Number Theory

Stillwell, J.
2003, XII, 256 p., Hardcover
ISBN: 978-0-387-95587-2