

1. Introduction

What is Number Theory?

In mathematics and its applications, we meet several different kinds of number.

Historically, and logically, we meet first the *natural numbers*, also known as the *positive integers*. These are the numbers used in counting. The set of such numbers is denoted by **N** (for natural numbers). We have

$$\mathbf{N} = \{ 1, 2, 3, 4, \dots \}.$$

There is good evidence that these were used over 30000 years ago.

Much later came the idea of *zero* and the *negative integers*. Together with the natural numbers, these give the set of *integers* or *whole numbers*. This is denoted by **Z** (for the German Zahlen, meaning *integers*). We have

$$\mathbf{Z} = \{ \dots, -3, -2, -1, 0, 1, 2, 3, \dots \}.$$

These numbers allow banking operations such as overdrafts. They were known in China some 3000 years ago, in India and Babylon some 2000 years ago. They did not reach Europe till about 1000 years ago, and were not in widespread use there until about 500 years ago.

Another kind of numbers arose from geometry. Early geometers knew how to divide a line segment *AB* into *n* equal parts, where *n* is a natural number. Even if *AB* has a length *m* which is a whole number, the equal pieces need not have *integral* length. For example, if *AB* has length 5 and *n* = 3, then each part has a length 5/3, which is *between* 1 and 2. In general, if *AB* has length *m*, then each of the *n* pieces has length given by the *quotient* or *fraction*, *m/n*. We can generalise this by allowing *m* to be negative, thus obtaining the set of rational numbers. This set is denoted by **Q** (for quotients). We have

$$\mathbf{Q} = \{ m/n : m \text{ in } \mathbf{Z}, n \text{ in } \mathbf{N} \}.$$

The instances with *m* positive were known to the geometers of Babylon, Egypt and Greece over 2500 years ago. The extension to general *m* came rather later when mathematicians began looking at numbers as solutions of equations. We will return to this idea shortly. It is important to notice that the notation is tricky. The same rational number may be written as a fraction in many ways. For example, 10/6 and 15/9 give the *same* rational (*i.e.* 5/3).

Geometers also discovered our final class of numbers. Followers of Pythagoras knew that square of unit side had a diagonal whose length was between 1 and 2, but *not* equal to any *rational* number. Each school of geometers knew that a circle of unit diameter had a circumference π which lay between $3 (= 21/7)$ and $22/7$. It was much later that it was proved that the value of π is *not* rational. If we imagine the number line, we can use geometry to mark any rational number provided we interpret *positive* and *negative* as *left* and *right*. Numbers such as π which are not rational (*i.e.* which are *irrational*) must correspond to *other* points on the line. The collection of *all* points on the number line is the set of *real numbers*. It is denoted by \mathbf{R} (for r real numbers).

Each of our number systems has two useful *arithmetic* operations. If we add or multiply two numbers from a single system, we obtain a third number, *of the same system*. For example, 2 and 3 belong to \mathbf{N} , and the sum, 5, and the product, 6, also lie in \mathbf{N} .

Observe that each of the number systems is contained in all later ones, *i.e.*

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R}.$$

Note that the *integer* m may be written as $m/1$, so is a member of \mathbf{Q} . Also, each member of \mathbf{Q} corresponds to a point on the number line, so gives a *real* number. What is not quite so obvious is that each number system contains numbers which are *not* in earlier ones. This can be understood by considering *equations*.

Consider the equation $x + 3 = 2$. This has no solution in \mathbf{N} since, for any positive x , the left hand side is *greater* than 2, so cannot equal 2. The solution $x = -1$ works in all larger systems. In a similar way, we can distinguish \mathbf{Z} from later systems by considering the equation $3x = 2$. Later we shall prove that the equation $x^2 = 2$ has no solution in \mathbf{Q} , though it has two solutions in \mathbf{R} , so \mathbf{Q} is a *smaller* set than \mathbf{R} .

Loosely, we can define Number Theory as the study of the *arithmetic* properties of \mathbf{Z} . We are in for some surprises.

Consider the equation $ax + by = c$, where a, b, c lie in \mathbf{N} . If we choose any y in \mathbf{Q} (or \mathbf{R}), then there is a *unique* value of x satisfying the equation. This may be computed as $(c - by)/a$. In \mathbf{Z} , the situation is quite different, and depends on the choice of a, b, c . We shall look at some examples.

Example 1 $4x + 5y = 9$.

This has an obvious solution $x = 1, y = 1$. A less obvious solution is $x = 6, y = -3$. There are in fact infinitely many solutions. Later, we shall show how they can *all* be found. On the other hand, there are values of x in \mathbf{Z} for which there is *no* corresponding y (in \mathbf{Z}). For example, if $x = 2$, we would have to put $y = 1/5$, which is *not* in \mathbf{Z} .

Finally, we observe that $x = y = 1$ is the *only* solution with x, y in \mathbf{N} . This can be deduced from the general solution mentioned above, but can be deduced *ad hoc*. If x, y are in \mathbf{N} , then $4x + 5y$ is at least $4 + 5 = 9$, and greater than 9 if either is greater than 1. Thus $x = y = 1$ is the only possibility.

Example 2 $4x - 5y = 9$.

This is very similar to Example 1. Indeed, the pair (x, y) is a solution of $4x - 5y = 9$ if and only if $(x, -y)$ is a solution of $4x + 5y = 9$. Now we have infinitely many solutions (x, y) with x and y in \mathbf{N} . The first three such pairs are $(6, 3), (11, 7)$ and $(16, 11)$.

Example 3 $4x + 5y = 1$.

We have obvious solutions $x = -1, y = 1$, and $x = 4, y = -3$. Again, there is an infinite family of solutions with x, y in \mathbf{Z} . But here, if x, y are in \mathbf{N} , then $4x + 5y$ is at least 9, so we have *no* solutions in \mathbf{N} .

Example 4 $4x + 6y = 9$.

Now there is *no* pair x, y in \mathbf{N} which give a solution. To see this, note that for integers x, y , $4x$ and $6y$ are *even* integers. Thus $4x+6y$ must be *even* – it *cannot* be equal to the *odd* integer 9.

Thus we see that the problem of solving an equation with variables in \mathbf{Q} or \mathbf{R} is quite different from the same problem with variables restricted to \mathbf{Z} or \mathbf{N} .

Definition A *diophantine equation* in variables x, y, z, \dots is an equation of the form

$$a_1T_1 + a_2T_2 + a_3T_3 + \dots = b,$$

where b and each a_i is in \mathbf{Z} , and each T_i is a product of powers of the variables.

For example, $3x^4 + xy^3 - 7x^3z + 12xyzt = 1$ is a diophantine equation in x, y, z, t .

Obviously, the equations in Examples 1-4 are diophantine equations in x, y .

Now we can formally define Number Theory as the study of diophantine equations with variables restricted to \mathbf{Z} .

We ask three questions about any diophantine equation :

- (1) Does the equation have *any* solutions with variables in \mathbf{Z} ?
- (2) Does the equation have *infinitely many* solutions with variables in \mathbf{Z} ?
- (3) Can we effectively find *all* the solutions?

We shall answer all three questions for equations of the form $ax + by = c$.

Not all problems are so easy to solve. About 400 years ago, Fermat discussed the equation

$$x^n + y^n = z^n.$$

This is a diophantine equation provided n is in \mathbf{N} . For $n = 2$, there are *many* solutions. The simplest example is $x = 3, y = 4, z = 5$. Indeed, Fermat himself showed how to find the entire family of solutions. He then made his famous conjecture :

for $n > 2$, the diophantine equation $x^n + y^n = z^n$ has *no* solutions in \mathbf{Z} with x, y, z non-zero.

It was only twenty years ago that the mathematician Wiles proved the truth of the conjecture. Understanding his proof requires thorough knowledge of many difficult areas of Number Theory.

What is cryptography?

In everyday language, the words *code* and *cipher* mean the same. Here, we find it useful to distinguish them.

A *code* is a method of transforming a message into a form suitable for transmission across a given channel of communication.

A *cipher* (also called a *cryptosystem*) is a method of transforming a message so that it can be understood only by the intended recipient.

The word *cryptography* comes from Greek words meaning *disguised writing*.

Cryptography used to be associated essentially with the military and espionage, but it is now seen as underpinning e-commerce. A buyer sends bank or credit card details, but does not want these to be understood by anyone other than the seller.

A familiar example of a code is the *ASCII code*, used to send text messages (strings of upper and lower case letters, digits, punctuation marks, spaces) over a *binary* channel (*i.e.* one which uses only the symbols 0 and 1 – the *bits*). Each text symbol is assigned an integer in the range 0-255. Each such integer is then encoded as an 8-bit word. Encoding and decoding are performed using *look-up tables* – lists showing each text symbol and the corresponding binary word. For example

A = 01000001,
B = 01000010,
a = 01100001.

An earlier example is the *Morse code*. Before telephony, messages were sent by telegraph. This uses only short pulses (*dots*), long pulses (*dashes*) and spaces. Each text symbol is encoded as a sequence of pulses. For example

E = dot,
S = dot, dot, dot
O = dash, dash, dash.

The sequences are separated by pauses. Otherwise, we could not distinguish between S and the string EEE. These systems *could* have used something like ASCII – just using dot and dash. Morse code speeds things up by sending common letters such as E in a short burst, and uncommon symbols such as Q in longer bursts.

Overall, a *code* transforms a message in one set of symbols (alphabet) into a message using a different set (another alphabet). To communicate, both sender and receiver must know the code. An outsider knowing the standard codes can understand an intercepted message. Thus, a code does not *disguise* the message in any useful way.

A cipher is used to disguise information. The message to be sent is called the *plaintext*. The message is sent as a transformed string (over the same alphabet). This is the *ciphertext*. The process of turning the plaintext into ciphertext is *encryption*. Of course, it must be possible for the intended recipient to recover the plaintext from the ciphertext. This process is *decryption*. Together, the processes of encryption and decryption form a *cryptosystem*.

The simplest kind of cryptosystem is a *substitution cryptosystem*. Each instance of a given symbol in the plaintext is replaced by a chosen symbol in the ciphertext. Different symbols in the plaintext must be replaced by different symbols in the ciphertext, otherwise decryption would be impossible. For example, if each of A and B in plaintext were replaced by C in ciphertext, then we could not decrypt C unambiguously.

For such systems, encryption and decryption are achieved by a look-up table. Each symbol of the plaintext is replaced by the symbol below it to get the ciphertext. For decryption, a symbol in the ciphertext is replaced by the symbol above it.

As an example, the following table gives a cryptosystem over the alphabet A,B, ...,Z.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B

In this cryptosystem, plaintext CAT encrypts as ECV. Ciphertext FQI decrypts as DOG.

The example is of a type known as a *shift cryptosystem*. Here, each symbol is replaced by that 2 steps further on, with a slight fudge at the end. We may replace 2 by any integer in the range 1, ..., 25, so we have 25 shift cryptosystems.

Shift cryptosystems are *very* insecure. There are only 25 possible shifts. To decrypt a shift forward by k steps, we shift *back* by k steps (equivalently, shift *forward* by $26-k$ steps). We can try each value of k in turn. 24 will produce gibberish. The correct value gives a stream of English words. This last can be recognised by a spellchecker.

The case of a general substitution cryptosystem is quite different. Such a system has a look-up table with the symbols A, ...,Z on top, and the symbols A, ...,Z *in some order* on the bottom. Now, there are $26! = 4 \times 10^{26}$ possible arrangements, and hence this many cryptosystems. If we exclude any where a symbol is encrypted as itself then some tricky mathematics shows that there are about 1.4×10^{26} systems. Now trial and error is useless. If a computer can test one billion cases per second, then it will take about five billion years to check them all! Such systems *seem* secure.

Suppose we have access to a *large amount* of ciphertext. If the symbols occur at random, then each symbol ought to occur $1/26 = 3.9\%$ of the time. In written English, the symbols do *not* occur randomly. In fact E occurs about 12.7% of the time. The next most common, T, occurs 9.6% of the time. Others occur less than 7%. Thus, from a large body of ciphertext, we can deduce the encrypted versions of E and T. At the other end of the scale, the symbols Q and Z are *very* infrequent. Thus, their encrypted versions can be identified. Once a number of symbols have been decrypted, the rest can be deduced since the plaintext consists of English words. This process is called *frequency analysis*.

Another defect of substitution cryptosystems is that, if the encryption process is known, then the decryption process is obvious – simply use the lookup table in reverse. It might be thought that any cryptosystem has this defect. After our study of Number Theory, we will be able to construct cryptosystems which are such that

- they are immune to frequency analysis,
- the encryption process does not lead to the decryption process (in reasonable time).