I

FACTORIZATION AND THE PRIMES

1. *The laws of arithmetic*

The object of the higher arithmetic is to discover and to establish general propositions concerning the natural numbers $1, 2, 3, \ldots$ of ordinary arithmetic. Examples of such propositions are the fundamental theorem (I.4)* that *every natural number can be factorized into prime numbers in one and only one way*, and Lagrange's theorem (V.4) that *every natural number can be expressed as a sum of four or fewer perfect squares*. We are not concerned with numerical calculations, except as illustrative examples, nor are we much concerned with numerical curiosities except where they are relevant to general propositions.

We learn arithmetic experimentally in early childhood by playing with objects such as beads or marbles. We first learn addition by combining two sets of objects into a single set, and later we learn multiplication, in the form of repeated addition. Gradually we learn how to calculate with numbers, and we become familiar with the laws of arithmetic: laws which probably carry more conviction to our minds than any other propositions in the whole range of human knowledge.

The higher arithmetic is a deductive science, based on the laws of arithmetic which we all know, though we may never have seen them formulated in general terms. They can be expressed as follows.

* References in this form are to chapters and sections of chapters of this book.

Addition. Any two natural numbers a and b have a *sum*, denoted by $a + b$, which is itself a natural number. The operation of addition satisfies the two laws:

$$a + b = b + a \qquad \textit{(commutative law of addition)},$$
$$a + (b + c) = (a + b) + c \quad \textit{(associative law of addition)},$$

the brackets in the last formula serving to indicate the way in which the operations are carried out.

Multiplication. Any two natural numbers a and b have a *product*, denoted by $a \times b$ or ab, which is itself a natural number. The operation of multiplication satisfies the two laws

$$ab = ba \qquad \textit{(commutative law of multiplication)},$$
$$a(bc) = (ab)c \quad \textit{(associative law of multiplication)}.$$

There is also a law which involves operations both of addition and of multiplication:

$$a(b + c) = ab + ac \quad \textit{(the distributive law)}.$$

Order. If a and b are any two natural numbers, then either a is equal to b or a is *less than* b or b is *less than* a, and of these three possibilities exactly one must occur. The statement that a is less than b is expressed symbolically by $a < b$, and when this is the case we also say that b is greater than a, expressed by $b > a$. The fundamental law governing this notion of order is that

$$\textit{if} \quad a < b \quad \textit{and} \quad b < c \quad \textit{then} \quad a < c.$$

There are also two other laws which connect the notion of order with the operations of addition and multiplication. They are that

$$\textit{if} \quad a < b \quad \textit{then} \quad a + c < b + c \quad \textit{and} \quad ac < bc$$

for any natural number c.

Cancellation. There are two laws of cancellation which, though they follow logically from the laws of order which have just been stated, are important enough to be formulated explicitly. The first is that

$$\textit{if} \quad a + x = a + y \quad \textit{then} \quad x = y.$$

This follows from the fact that if $x < y$ then $a + x < a + y$, which is contrary to the hypothesis, and similarly it is impossible that $y < x$, and therefore $x = y$. In the same way we get the second law of cancellation, which states that

$$\textit{if} \quad ax = ay \quad \textit{then} \quad x = y.$$

Subtraction. To subtract a number b from a number a means to find, if possible, a number x such that $b + x = a$. The possibility of subtraction is related to the notion of order by the law that *b can be subtracted from a if and only if b is less than a*. It follows from the first cancellation law that if subtraction is possible, the resulting number is unique; for if $b + x = a$ and $b + y = a$ we get $x = y$. The result of subtracting b from a is denoted by $a - b$. Rules for operating with the minus sign, such as $a - (b - c) = a - b + c$, follow from the definition of subtraction and the commutative and associative laws of addition.

Division. To divide a number a by a number b means to find, if possible, a number x such that $bx = a$. If such a number exists it is denoted by $\frac{a}{b}$ or a/b. It follows from the second cancellation law that if division is possible the resulting number is unique.

All the laws set out above become more or less obvious when one gives addition and multiplication their primitive meanings as operations on sets of objects. For example, the commutative law of multiplication becomes obvious when one thinks of objects arranged in a rectangular pattern with a rows and b columns (fig. 1); the total number of objects is ab and is also ba. The distributive law becomes obvious when one considers the arrangement of objects indicated in fig. 2; there are $a(b + c)$ objects altogether and these are made up of ab objects together with ac more objects. Rather less obvious, perhaps, is the associative law of multiplication, which asserts that $a(bc) = (ab)c$. To make this apparent, consider the same rectangle as in fig. 1, but replace each object by the number c. Then the sum of all the numbers in any one row is bc, and as there are a rows the total sum is $a(bc)$. On the other hand, there are altogether ab numbers each of which is c, and therefore the total sum is $(ab)c$. It follows that $a(bc) = (ab)c$, as stated.

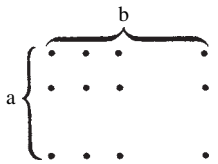

Fig. 1

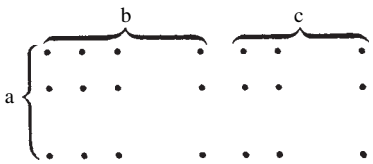

Fig. 2

The laws of arithmetic, supplemented by the principle of induction (which we shall discuss in the next section), form the basis for the logical development of the theory of numbers. They allow us to prove general theorems about the natural numbers without it being necessary to go back to the primitive meanings of the numbers and of the operations carried out

on them. Some quite advanced results in the theory of numbers, it is true, are most easily proved by counting the same collection of things in two different ways, but there are not very many such.

Although the laws of arithmetic form the logical basis for the theory of numbers (as indeed they do for most of mathematics), it would be extremely tedious to refer back to them for each step of every argument, and we shall in fact assume that the reader already has some knowledge of elementary mathematics. We have set out the laws in detail in order to show where the subject really begins.

We conclude this section by discussing briefly the relationship between the system of natural numbers and two other number-systems that are important in the higher arithmetic and in mathematics generally, namely the *system of all integers* and the *system of all rational numbers*.

The operations of addition and multiplication can always be carried out, but those of subtraction and division cannot always be carried out within the natural number system. It is to overcome the limited possibility of subtraction that there have been introduced into mathematics the number 0 and the negative integers $-1, -2, \ldots$. These, together with the natural numbers, form the system of all integers:

$$\ldots, -2, -1, 0, 1, 2, \ldots,$$

within which subtraction is always possible, with a unique result. One learns in elementary algebra how to define multiplication in this extended number-system, by the 'rule of signs', in such a way that the laws of arithmetic governing addition and multiplication remain valid. The notion of order also extends in such a way that the laws governing it remain valid, with one exception: the law that if $a < b$ then $ac < bc$ remains true only if c is positive. This involves an alteration in the second cancellation law, which is only true in the extended system if the factor cancelled is not 0:

$$\textit{if} \quad ax = ay \quad \textit{then} \quad x = y, \quad \textit{provided that} \quad a \neq 0.$$

Thus the integers (positive, negative and zero) satisfy the same laws of arithmetic as the natural numbers except that subtraction is now always possible, and that the law of order and the second cancellation law are modified as just stated. The natural numbers can now be described as the *positive integers*.

Let us return to the natural numbers. As we all know, it is not always possible to divide one natural number by another, with a result which is itself a natural number. If it is possible to divide a natural number b by a natural number a within the system, we say that a is a *factor* or *divisor* of b, or that b is a *multiple* of a. All these express the same thing. As illustrations

of the definition, we note that 1 is a factor of every number, and that a is itself a factor of a (the quotient being 1). As another illustration, we observe that the numbers divisible by 2 are the even numbers $2, 4, 6, \ldots$, and those not divisible by 2 are the odd numbers $1, 3, 5, \ldots$.

The notion of divisibility is one that is peculiar to the theory of numbers, and to a few other branches of mathematics that are closely related to the theory of numbers. In this first chapter we shall consider various questions concerning divisibility which arise directly out of the definition. For the moment, we merely note a few obvious facts.

(i) *If a divides b then $a \leqq b$* (that is, a is either less than or equal to b). For $b = ax$, so that $b - a = a(x - 1)$, and here $x - 1$ is either 0 or a natural number.

(ii) *If a divides b and b divides c then a divides c*. For $b = ax$ and $c = by$, whence $c = a(xy)$, where x and y denote natural numbers.

(iii) *If two numbers b and c are both divisible by a, then $b + c$ and $b - c$ (if $c < b$) are also divisible by a.* For $b = ax$ and $c = ay$, whence

$$b + c = a(x + y) \text{ and } b - c = a(x - y).$$

There is no need to impose the restriction that $b > c$ when considering $b - c$ in the last proposition, if we extend the notion of divisibility to the integers as a whole in the obvious way: an integer b is said to be divisible by a natural number a if the quotient $\frac{b}{a}$ is an integer. Thus a negative integer $-b$ is divisible by a if and only if b is divisible by a. Note that 0 is divisible by every natural number, since the quotient is the integer 0.

(iv) *If two integers b and c are both divisible by the natural number a, then every integer that is expressible in the form $ub + vc$, where u and v are integers, is also divisible by a.* For $b = ax$ and $c = ay$, whence $ub + vc = (ux + vy)a$. This result includes those stated in (iii) as special cases; if we take u and v to be 1 we get $b + c$, and if we take u to be 1 and v to be -1 we get $b - c$.

Just as the limitation on the possibility of subtraction can be removed by enlarging the natural number system through the introduction of 0 and the negative integers, so also the limitation on the possibility of division can be removed by enlarging the natural number system through the introduction of all positive fractions, that is, all fractions $\frac{a}{b}$, where a and b are natural numbers. If both methods of extension are combined, we get the *system of rational numbers*, comprising all integers and all fractions, both positive and negative. In this system of numbers, all four operations

of arithmetic—addition, multiplication, subtraction and division—can be carried out without limitation, except that division by zero is necessarily excluded.

The main concern of the theory of numbers is with the natural numbers. But it is often convenient to work in the system of all integers or in the system of rational numbers. It is, of course, important that the reader, when following any particular train of reasoning, should note carefully what kinds of numbers are represented by the various symbols.

2. *Proof by induction*

Most of the propositions of the theory of numbers make some assertion about every natural number; for example Lagrange's theorem asserts that every natural number is representable as the sum of at most four squares. How can we prove that an assertion is true for *every natural number*? There are, of course, some assertions that follow directly from the laws of arithmetic, as for instance algebraic identities like

$$(n+1)^2 = n^2 + 2n + 1.$$

But the more interesting and more genuinely arithmetical propositions are not of this simple kind.

It is plain that we can never prove a general proposition by verifying that it is true when the number in question is 1 or 2 or 3, and so on, because we cannot carry out infinitely many verifications. Even if we verify that a proposition is true for every number up to a million, or a million million, we are no nearer to establishing that it is true always. In fact it has sometimes happened that propositions in the theory of numbers, suggested by extensive numerical evidence, have proved to be wide of the truth.

It may be, however, that we can find a *general argument* by which we can prove that *if* the proposition in question is true for all the numbers

$$1, 2, 3, \ldots, n-1,$$

then it is true for the next number, n. If we have such an argument, then the fact that the proposition is true for the number 1 will imply that it is true for the next number, 2; and then the fact that it is true for the numbers 1 and 2 will imply that it is true for the number 3, and so on indefinitely. The proposition will therefore be true for every natural number if it is true for the number 1.

This is the principle of proof by induction. The principle relates to propositions which assert that something is true for every natural number, and in order to apply the principle we need to prove two things: first, that the

assertion in question is true for the number 1, and secondly that *if* the assertion is true for each of the numbers $1, 2, 3, \ldots, n-1$ preceding any number n, *then* it is true for the number n. Under these circumstances we conclude that the proposition is true for every natural number.

A simple example will illustrate the principle. Suppose we examine the sum $1 + 3 + 5 + \cdots$ of the successive odd numbers, up to any particular one. We may notice that

$$1 = 1^2, 1 + 3 = 2^2, 1 + 3 + 5 = 3^2, 1 + 3 + 5 + 7 = 4^2,$$

and so on. This suggests the general proposition that *for every natural number n, the sum of the first n odd numbers is* n^2. Let us prove this general proposition by induction. It is certainly true when n is 1. Now we have to prove that the result is true for any number n, and by the principle of induction we are entitled to suppose that it is already known to be true for any number less than n. In particular, therefore, we are entitled to suppose that we already know that the sum of the first $n - 1$ odd numbers is $(n - 1)^2$. The sum of the first n odd numbers is obtained from this by adding the nth odd number, which is $2n - 1$. So the sum of the first n odd numbers is

$$(n - 1)^2 + (2n - 1),$$

which is in fact n^2. This proves the proposition generally.

Proofs by induction are sometimes puzzling to the inexperienced, who are liable to complain that 'you are assuming the proposition that is to be proved'. The fact is, of course, that a proposition of the kind now under consideration is a proposition with an infinity of cases, one for each of the natural numbers $1, 2, 3, \ldots$; and all that the principle of induction allows us to do is to suppose, when proving any one case, that the preceding cases have already been settled.

Some care is called for in expressing a proof by induction in a form which will not cause confusion. In the example above, the proposition in question was that *the sum of the first n odd numbers is* n^2. Here n is any one of the natural numbers, and, of course, the statement means just the same if we change n into any other symbol, provided we use the same symbol in the two places where it occurs. But once we have embarked on the proof, n becomes a particular number, and we are then in danger of using the same symbol in two senses, and even of writing such nonsense as 'the proposition is true when n is $n-1$'. The proper course is to use different symbols where necessary.

From a commonsense point of view, nothing can be more obvious than the validity of proof by induction. Nevertheless it is possible to debate whether the principle is in the nature of a *definition* or a *postulate* or an *act*

of faith. What seems at any rate plain is that the principle of induction is essentially a statement of the rule by which we enumerate the natural numbers in order: having enumerated the numbers $1, 2, \ldots, n-1$ we continue the enumeration with the next number n. Thus the principle is in effect an explanation of what is meant by the words 'and so on', which must occur whenever we attempt to enumerate the natural numbers.

3. *Prime numbers*

Obviously any natural number a is divisible by 1 (the quotient being a) and by a (the quotient being 1). A factor of a other than 1 or a is called a *proper* factor. We all know that there are some numbers which have *no* proper factors, and these are called prime numbers, or *primes*. The first few primes are

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

Whether 1 should be counted as a prime or not is a matter of convention, but it is simpler (as we shall see later) not to count 1 as a prime.

A number which is neither 1 nor a prime is said to be *composite*; such a number is representable as the product of two numbers, each greater than 1. It is well known that by continued factorization one can eventually express any composite number as a product of primes, some of which may of course be repeated. For example, if we take the number 666, this has the obvious factor 2, and we get $666 = 2 \times 333$. Now 333 has the obvious factor 3, and $333 = 3 \times 111$. Again 111 has the factor 3, and $111 = 3 \times 37$. Hence

$$666 = 2 \times 3 \times 3 \times 37,$$

and this is a representation of the composite number 666 as a product of primes. The general proposition is that any composite number is representable as a product of primes. Or, what comes to the same thing, *any number greater than* 1 *is either a prime or is expressible as a product of primes*.

To prove this general proposition, we use the method of induction. In proving the statement for a number n, we are entitled to assume that it has already been proved for any number less than n. If n is a prime, there is nothing to prove. If n is composite, it can be represented as ab, where a and b are both greater than 1 and less than n. We know that a and b are either primes or are expressible as products of primes, and on substituting for them we get n expressed as a product of primes. This proof is indeed so simple that the reader may think it quite superfluous. But the next general proposition on factorization into primes will not be so easily proved.

The series 2, 3, 5, 7, ... of primes has always exercised human curiosity, and later we shall mention some of the results that are known about it. For the moment, we content ourselves with proving, following Euclid (Book IX, Prop. 20), that *the series of primes never comes to an end*. His proof is a model of simplicity and elegance. Let 2, 3, 5, ..., P be the series of primes up to a particular prime P. Consider the number obtained by multiplying all these primes together, and then adding 1, that is

$$N = 2 \times 3 \times 5 \times \cdots \times P + 1.$$

This number cannot be divisible by 2, for then both the numbers N and $2 \times 3 \times 5 \times \cdots \times P$ would be divisible by 2, and therefore their difference would be divisible by 2. This difference is 1, and is not divisible by 2. In the same way, we see that N cannot be divisible by 3 or by 5 or by any of the primes up to and including P. On the other hand, N is divisible by *some* prime (namely N itself if N is a prime, or any prime factor of N if N is composite). Hence there exists a prime which is different from any of the primes 2, 3, 5, ..., P, and so is greater than P. Consequently the series of primes never comes to an end.

4. *The fundamental theorem of arithmetic*

It was proved in the preceding section that any composite number is expressible as a product of primes. As an illustration, we factorized 666 and obtained

$$666 = 2 \times 3 \times 3 \times 37.$$

A question of fundamental importance now suggests itself. Is such a factorization into primes possible in more than one way? (It is to be understood, of course, that two representations which differ merely in the order of the factors are to be considered as the same, e.g. the representation $3 \times 2 \times 37 \times 3$ is to be considered the same as that printed above.) Can we conceive that 666, for example, has some other representation as a product of primes? The reader who has no knowledge of the theory of numbers will probably have a strong feeling that no other representation is possible, but he will not find it a very easy matter to construct a satisfactory general proof.

It is convenient to express the proposition in a form in which it applies to all natural numbers, and not only to composite numbers. If a number is itself a prime, we make the convention that it is to be regarded as a 'product' of primes, where the 'product' has only one factor, namely the number itself. We can go even a stage further, and regard the number 1 as an 'empty'

product of primes, making the convention that the value of an empty product is deemed to be 1. This is a convention which is useful not only here but throughout mathematics, since it permits the inclusion in general theorems of special cases which would otherwise have to be excluded, or provided for by a more complicated enunciation.

With these conventions, the general proposition is that *any natural number can be represented in one and only one way as a product of primes*. This is the so-called *fundamental theorem of arithmetic*, and its history is strangely obscure. It does not figure in Euclid's *Elements*, though some of the arithmetical propositions in Book VII of the *Elements* are almost equivalent to it. Nor is it stated explicitly even in Legendre's *Essai sur la théorie des nombres* of 1798. The first clear statement and proof seem to have been given by Gauss in his famous *Disquisitiones Arithmeticae* of 1801. Perhaps the omission of the theorem from Euclid explains why it is passed over without explanation in many schoolbooks. One of them (still in use) describes it as a 'law of thought', which it certainly is not.

We now give a direct proof of the uniqueness of factorization into primes. Later (in §7) we shall give another proof, which will be entirely independent of the present one.

First there is a preliminary remark to be made. *If* the factorization of a particular number m into primes is unique, each prime factor of m must occur in that factorization. For if p is any prime which divides m, we have $m = pm'$ where m' is some other number, and if we now factorize m' into primes we obtain a factorization of m into primes by simply putting on the additional factor p. Since there is supposed to be only one factorization of m into primes, p must occur in it.

We prove the uniqueness of factorization by induction. This requires us to prove it for any number n, on the assumption that it is already established for all numbers less than n. If n is itself a prime, there is nothing to prove. Suppose, then, that n is composite, and has two different representations as products of primes, say

$$n = pqr \ldots = p'q'r' \ldots,$$

where $p, q, r, \ldots$ and $p', q', r', \ldots$ are all primes. The same prime cannot occur in both representations, for if it did we could cancel it and get two different representations of a smaller number, which is contrary to the inductive hypothesis.

We can suppose without loss of generality that p is the least of the primes occurring in the first factorization. Since n is composite, there is at least one prime besides p in the factorization, and therefore $n \geqq p^2$. Similarly