INTRODUCTION

Everybody, or nearly everybody, knows the Fundamental Theorem of Arithmetic, even if not by name. It goes back to the ancient Greek mathematicians. Start with a whole number, preferably positive, say 60. Try factorizing 60 into smaller and smaller factors until you cannot do it any further, and you end up with $60 = 2 \times 2 \times 3 \times 5$. The factors 2, 2, 3, 5 are 'irreducible' in that they cannot be further factorized. Another name for them is that they are 'prime'. Moreover, this is just about the only way of factorizing 60 into irreducibles. You could, for example, write $60 = (-2) \times 2 \times (-3) \times 5$ if you insisted, or you could change the order of the factors, but most reasonable people would say that these factorizations are no different from the previous one. Thus there is essentially one and only one way of factorizing 60 into irreducible (or prime) factors, and this is true for all whole numbers. This is the Fundamental Theorem of Arithmetic.

The same sort of thing happens with polynomials. Take the polynomial $x^3 + x^2 + x + 1$, try factorizing as far as you can, and you end up with

$$x^3 + x^2 + x + 1 = (x + 1)(x^2 + 1).$$

At least, that is what you end up with if you insist that all the coefficients are real numbers. If you are brave enough to allow complex coefficients, you can get further, and write

$$x^3 + x^2 + x + 1 = (x + 1)(x + \mathrm{i})(x - \mathrm{i}).$$

Another fundamental theorem, called this time the Fundamental Theorem of Algebra, tells us that a polynomial with complex coefficients can be factorized into linear factors; and this can be done in essentially only one way. You could, for example, write

$$x^3 + x^2 + x + 1 = (\mathrm{i}x + \mathrm{i})(-\mathrm{i}x + 1)(x - \mathrm{i}),$$

but most reasonable people would say that this is the same factorization as the previous one; the first factor has been multiplied by i and the second by $-$i, which has little effect overall. This famous theorem was first proved by the great Gauss in his doctoral thesis in 1799, although the result was known before then. It is in fact not a theorem in algebra at all, but one in analysis, and Gauss's proofs (for he gave more than one) would not stand the test of today's higher standards of rigour. If the polynomial you start with has real coefficients and you insist that the factors have real coefficients, then the polynomial can only be factorized into irreducible linear and quadratic factors.

Here are two very different situations where the same thing happens, firstly with integers and secondly with polynomials. The questions naturally arise, does it happen in other situations and does it always happen? The answers are 'yes' and 'no', respectively. We shall need to set up a suitable algebraic framework in which to consider these questions, and this brings us to the idea of a 'ring'. Roughly (very roughly) speaking, to get the idea of a ring, just think what properties whole numbers (integers) and polynomials have in common, write them down and say that anything that satisfies these properties is a ring. In particular, you can add, subtract and multiply in a ring subject to some rather obvious-looking rules. The whole point of moving into this abstract setting rather than staying with the more familiar whole numbers, or polynomials, is that you can deal with both situations, and hopefully many more, at the same time.

But why bother? Isn't it just abstraction for abstraction's sake, the curse of today's unfortunate student? Hopefully not! In the first place, a number of results about whole numbers come out of this abstraction in a very elegant way. But, more importantly, factorization in this more abstract setting was forced upon mathematicians by perhaps the most famous unsolved problem of all in mathematics, Fermat's Last 'Theorem' (in inverted commas because no one has managed to prove or disprove it, yet). We all know that $3^2+4^2=5^2$, so that there are positive integers x, y, z such that $x^2+y^2=z^2$. The French mathematician Pierre Fermat proposed in about 1637 that, when $n>2$, there are no positive integers x, y, z such that

$x^n + y^n = z^n$. He wrote in the margin of his copy of Bachet's translation of Diophantus' *Arithmetica* that he had discovered the most remarkable short proof of this result, but that the margin was too small to contain it. Ever since, mathematicians, professional and amateur alike, have tried and failed to prove it. Gauss gave a proof for $n = 3$ which involves the idea of factorization in a ring other than the two mentioned so far. In 1843, Kummer gave what he thought was a proof of the general result, but Dirichlet pointed out to him that he had assumed that unique factorization held in a particular setting in which it did not in fact hold. It was this that brought out the importance of factorization in general, and led Dedekind and others to restore unique factorization not for single elements any more but for whole sets of elements called 'ideals'. But this takes us beyond the scope of this little book. We can only hope that this book will whet readers' appetites for further study.

Factorization of integers into their prime factors has recently come into prominence in a surprising way. Security-conscious governments are naturally concerned to be able to pass messages without their being intercepted, and this is done in code, hopefully in a code which cannot be broken by an 'enemy'. Such a coding technique is provided by the so-called 'Public-Key Cryptography'. The receiver of the message starts with two (or more) large prime numbers, say P, Q, having of the order of 50 to 100 digits. He gives the sender of the message only the product PQ, which the sender uses to encode the given message. The message can only be decoded when the individual primes P, Q are known. Herein lies the difficulty in breaking the code. The time required to factorize a large number into its prime factors increases exponentially with the size of the number, unlike that required to test that a given integer is prime, which only increases linearly with the size of the number. Thus, even in these days of very large computers, it is not practicable to extract the individual factors P, Q just from a knowledge of their product PQ. Thus, at the moment, and unless you are very unlucky (i.e. your 'enemy' is very lucky), your code is secure. This puts factorization of numbers in the forefront of research by today's computer scientists and makes it of interest to more than undergraduates in mathematics! For a popular article on this subject, reprinted from *The*

Guardian, we refer readers to 'Prime Numbers and Secret Codes' by Keith Devlin, in *Mathematical Spectrum*, where references to research articles may be found†.

But now it is time to come down from these dizzy heights and get down to details.

† Added in proof. It has just been announced that H. W. Lenstra, University of Amsterdam, has devised a technique which makes the factorization of large numbers more possible, so maybe security services the world over will have to think again!

Part 1

Rings

1.1

Introduction

The first mathematics that most of us met concerned whole numbers or integers,

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots.$$

We learned how to add, subtract and multiply them. Later on, we thought how to divide them, but that introduces fractions so we will lay it aside for the moment. With increased sophistication, we dealt with polynomials such as $x^2 + 2$, $x^3 - 3x^2 + 1$, and added, subtracted and multiplied these as well. Both of these systems satisfy the same simple laws for addition, subtraction and multiplication. In fact, these laws are usually used without thinking. However, if we are to consider other systems than these, we shall need to make a note of the properties that we need and, in theory at any rate, we should check that the properties are satisfied each time a new system is introduced. These properties are incorporated in the definition of a ring, first formally given by A. A. Fraenkel in 1914.

1.2

Binary operations

First the idea of a binary operation on a set. Let S be a non-empty set. A *binary operation* on S is a rule whereby, given elements a, b of S (which could be the same element), there is defined a unique element of S, denoted variously by $a + b$ or ab or $a \cdot b$ or $a \circ b$ according to how the binary operation is denoted. For example, if $\mathbb{Z}$ denotes the set of integers (as it always does following the German word '*zahlen*',

meaning 'number'), then addition and multiplication are both binary operations on $\mathbb{Z}$ in that, given integers a, b, there are defined unique integers $a + b$, ab in $\mathbb{Z}$. If we are considering the binary operation $+$ on $\mathbb{Z}$, we sometimes but not always write $(\mathbb{Z}, +)$ to emphasize this, and $(\mathbb{Z}, \cdot)$ or $(\mathbb{Z}, \times)$ when we are considering multiplication on $\mathbb{Z}$. If we are considering both addition and multiplication on $\mathbb{Z}$, we may write $(\mathbb{Z}, +, \cdot)$. Generally, if we have a non-empty set S and a binary operation denoted by $+$ defined on S, we write $(S, +)$; and $(S, +, \cdot)$ will denote S with two binary operations on it, denoted by addition and multiplication. We emphasize that the elements of S may not be numbers and that the binary operations, although they may be denoted by addition and multiplication, may have nothing to do with addition and multiplication of numbers. Surprisingly, it is less confusing to stick to additive and multiplicative notation rather than to introduce such devices as $\circ$ and $*$ to denote a binary operation on a set.

This definition of a binary operation on S may leave you feeling a little dissatisfied or cheated. After all, what is a 'rule'? For the more sophisticated reader, a binary operation on S is a mapping from the Cartesian product $S \times S$ to S. The Cartesian product $S \times S$ consists of all ordered pairs (a, b), where a, b are elements of S (similar to the Cartesian coordinates of points in a plane), and the image of the pair (a, b) under the mapping is denoted by $a + b$ or ab or $a \circ b$ according to the way in which the binary operation is designated. But the sophisticated way of thinking of a binary operation is seldom the best, so we shall resort to the simple-minded way.

The examples which follow are all examples of well-known sets with two binary operations defined on them. As well as illustrating the notion of a binary operation, they will also serve to establish some standard notation for various sets.

Examples of sets with binary operations

(1) $(\mathbb{Z}, +, \cdot)$, where $\mathbb{Z}$ is the set of integers.

(2) $(\mathbb{Q}, +, \cdot)$, where $\mathbb{Q}$ is the set of rational numbers.

(3) $(\mathbb{R}, +, \cdot)$, where $\mathbb{R}$ is the set of real numbers.

(4) $(\mathbb{C}, +, \cdot)$, where $\mathbb{C}$ is the set of complex numbers.

(5) $(M_n(\mathbb{R}), +, \cdot)$, where $M_n(\mathbb{R})$ denotes the set of all $n \times n$ matrices with real entries, and $+, \cdot$ denote the usual matrix addition and multiplication.

(6) $(\mathbb{R}[x], +, \cdot)$, where $\mathbb{R}[x]$ denotes the set of all polynomials in x with real coefficients and $+, \cdot$ denote the usual addition and multiplication of polynomials.

(7) $(\mathbb{R}[[x]], +, \cdot)$, where $\mathbb{R}[[x]]$ denotes the set of all 'formal power series'. These are formal expressions of the form

$$a_0 + a_1 x + a_2 x^2 + \ldots,$$

where the coefficients $a_0, a_1, a_2, \ldots$ are all real numbers. Although this looks like an infinite sum, it is not, and no concept of convergence is involved, as would be the case if we were trying to add infinitely many things together. The formal expression really stands for the infinite sequence $(a_0, a_1, a_2, \ldots)$, and the powers of x are merely used to denote the positions in the sequence. It should be noted that

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n,$$

where the a_n and b_n are real numbers, if and only if $a_n = b_n$ for every n. Addition and multiplication on $\mathbb{R}[[x]]$ are defined by

$$\left(\sum_{n=0}^{\infty} a_n x^n\right) + \left(\sum_{n=0}^{\infty} b_n x^n\right) = \sum_{n=0}^{\infty} (a_n + b_n) x^n,$$

$$\left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=0}^{\infty} b_n x^n\right) = \sum_{n=0}^{\infty} (a_0 b_n + a_1 b_{n-1} + \ldots + a_n b_0) x^n.$$

These rules are similar to the rules for the addition and multiplication of polynomials. In fact, $\mathbb{R}[x]$ is a subset of $\mathbb{R}[[x]]$ in that a polynomial is just a power series with only finitely many non-zero coefficients.

(8) Instead of considering polynomials in a single variable x, we can have polynomials in n independent variables (usually called 'indeterminates') $x_1, \ldots, x_n$ with real coefficients (say). We denote the set of these by $\mathbb{R}[x_1, \ldots, x_n]$. Such a polynomial is a formal expression of the form

$$\sum a_{i_1 \ldots i_n} x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n},$$

where $i_1, \ldots, i_n$ are non-negative integers and $a_{i_1 \ldots i_n} \in \mathbb{R}$† with only

† The symbol $\in$ stands for 'belongs to'. Thus $a \in S$ means that a is an element of the set S. Also, $a \notin S$ means that a is not an element of S.

finitely many of the $a_{i_1 \ldots i_n}$ non-zero. (For instance, an example of an element of $\mathbb{R}[x_1, x_2, x_3]$ is $4x_1^2x_2^3x_3 - 6x_2x_3^4 + x_1^4$; this is also an element of the set $\mathbb{Z}[x_1, x_2, x_3]$ since its coefficients are actually integers.) Addition and multiplication can be defined on $\mathbb{R}[x_1, \ldots, x_n]$ by

$$\begin{aligned}
&(\sum a_{i_1 \ldots i_n} x_1^{i_1} \ldots x_n^{i_n}) + (\sum b_{i_1 \ldots i_n} x_1^{i_1} \ldots x_n^{i_n}) \\
&\quad = \sum (a_{i_1 \ldots i_n} + b_{i_1 \ldots i_n}) x_1^{i_1} \ldots x_n^{i_n}, \\
&(\sum a_{i_1 \ldots i_n} x_1^{i_1} \ldots x_n^{i_n})(\sum b_{i_1 \ldots i_n} x_1^{i_1} \ldots x_n^{i_n}) \\
&\quad = \sum \Big(\sum_{\substack{j_r + k_r = i_r \\ 1 \leqslant r \leqslant n}} a_{j_1 \ldots j_n} b_{k_1 \ldots k_n} \Big) x_1^{i_1} \ldots x_n^{i_n}.
\end{aligned}$$

Examples of these rather complicated looking rules would be

$$\begin{aligned}
&(4x_1^2x_2^3x_3 - 3x_2x_3^4 + x_1^4) + (x_1^2x_2^3x_3 + 3x_2x_3) \\
&\quad = 5x_1^2x_2^3x_3 - 3x_2x_3^4 + x_1^4 + 3x_2x_3, \\
&(4x_1^2x_2^3x_3 - 3x_2x_3 + x_1^4)(x_1^2x_2^3x_3 + 3x_2x_3) \\
&\quad = 4x_1^4x_2^6x_3^2 + (12-3)x_1^2x_2^4x_3^2 - 9x_2^2x_3^2 \\
&\qquad + x_1^6x_2^3x_3 + 3x_1^4x_2x_3.
\end{aligned}$$

(9) $(C[a, b], +, \cdot)$, where $C[a, b]$ denotes the set of all continuous real-valued functions defined on a closed interval $[a, b]$ and $+, \cdot$ are defined pointwise, i.e.

$$(f+g)(x) = f(x) + g(x), (fg)(x) = f(x)g(x).$$

Exercise 1.2 Think of ten other examples of sets with one or two binary operations defined on them. Try to make them as different as you can.

1.3 Definition of a ring

As we list the axioms that must be satisfied by a ring, it is suggested that readers bear in mind the prototype examples of the integers and polynomials. In fact, all the examples listed in Section 1.2 of sets with two binary operations will satisfy the axioms and so are rings.

Definition 1.3.1 A 'ring' is a non-empty set R which satisfies the following axioms:

(1) R has a binary operation denoted by $+$ defined on it;

(2) addition is associative, i.e.

$$a+(b+c)=(a+b)+c \text{ for all } a, b, c \in R$$

(so that we can write $a+b+c$ without brackets);

(3) addition is commutative, i.e.

$$a+b=b+a \text{ for all } a, b \in R;$$

(4) there is an element denoted by 0 in R such that

$$0+a=a \text{ for all } a \in R$$

(there is only one such element because, if 0_1, 0_2 are two such, then $0_1 = 0_1+0_2=0_2$ and they are the same – we call 0 the *zero element* of R);

(5) for every $a \in R$, there exists an element $-a \in R$ such that

$$(-a)+a=0$$

(there is only one such element for each a, because if $b+a=0$ and $c+a=0$, then

$$b=0+b=(c+a)+b=c+(a+b)=c+0=c;$$

we call $-a$ the *negative* of a);

(6) R has a binary operation denoted by multiplication defined on it;

(7) multiplication is associative, i.e.

$$a(bc)=(ab)c \text{ for all } a, b, c \in R;$$

(8) multiplication is left and right distributive over addition, i.e.

$$a(b+c)=ab+ac, (a+b)c=ac+bc \text{ for all } a, b, c \in R;$$

(9) there is an element denoted by 1 in R such that $1 \neq 0$ and

$$1 \cdot a=a \cdot 1=a \text{ for all } a \in R$$

(as for the zero element, there is only one such element, and it is called the *identity element* of R).

Axioms 1–5 may be summarized by saying that R is an abelian group under addition.

Axiom 9 is not imposed by all authors. Thus, for example, the even integers with the usual addition and multiplication would form a ring without an identity element. However, we shall insist that our rings possess identity elements and that $1 \neq 0$. It may be asked how 1 and 0 could be equal. If we take a single-element set $\{x\}$ and define addition and multiplication on it by $x+x=x$, $xx=x$, then we obtain a 'ring' in which $0=x=1$. Such a ring, with only one element, is called a *trivial ring*. In fact, to anticipate Theorem 1.3.2, if $1=0$ in a ring, then $a=a\cdot 1=a\cdot 0=0$ for all a, and such a ring must be a trivial ring. Thus trivial rings provide the only situation in which $1=0$. Thus trivial rings are excluded by our Axiom 9.

The next result includes some elementary consequences of the definition which are usually used without thinking.

Theorem 1.3.2 Let R be a ring. Then

(1) $0a=a0=0$ for all $a\in R$,
(2) $(-a)b=a(-b)=-(ab)$ for all $a, b\in R$,
(3) $(-a)(-b)=ab$ for all $a, b\in R$.

Proof (1) $0a+0a=(0+0)a=0a$†, so that

$$\begin{aligned} -(0a)+(0a+0a) &= -(0a)+0a, \\ (-(0a)+0a)+0a &= 0, \\ 0+0a &= 0, \\ 0a &= 0. \end{aligned}$$

A similar argument shows that $a0=0$.

(2) $(-a)b+ab=((-a)+a)b=0b=0$ by (1), so that $(-a)b=-(ab)$. A similar argument shows that $a(-b)=-(ab)$.

(3) $$\begin{aligned} (-a)(-b) &= -(a(-b)) \text{ by (2)} \\ &= -(-(ab)) \text{ by (2)} \\ &= ab \text{ by Axiom 5.} \quad \square \end{aligned}$$

† The usual conventions concerning addition and multiplication apply. Thus $0a+0a$ means $(0a)+(0a)$.