2 Rings

Rings and subrings

This chapter is about rings, which are abstract systems in which addition and multiplication are defined. The prototype is the ring $\mathbb{Z}$ of integers, with the ordinary arithmetic operations. This is the example to which we keep coming back: we are interested in seeing how far the familiar properties of integers (as outlined in Chapter 1) can be extended.

2.1 Introduction. A **ring** is a set with two binary operations called **addition** and **multiplication**. We use the same notation for these operations in a general ring as in the integers: addition is represented by an infix $+$, and multiplication by either juxtaposition or an infix $\cdot$. (That is, we denote the sum of a and b by $a+b$, and their product by either ab or $a \cdot b$.)

A ring is defined by a list of axioms, which follows. These are divided into three groups, involving addition, multiplication, and both operations, respectively. These are meant to be familiar properties of $\mathbb{Z}$. I will assume without proof that they all hold in $\mathbb{Z}$.

Axioms for addition
(A0) (**Closure law**): For all $a, b \in R$, $a + b \in R$.
(A1) (**Associative law**): $a + (b + c) = (a + b) + c$ for all $a, b, c \in R$.
(A2) (**Zero law**): There exists $0 \in R$ such that $a + 0 = 0 + a = a$ for all $a \in R$.
(A3) (**Inverse law**): For all $a \in R$, there exists $b \in R$ with $a + b = b + a = 0$.
(A4) (**Commutative law**): $a + b = b + a$ for all $a, b \in R$.
Axioms for multiplication
(M0) (**Closure law**): For all $a, b \in R$, $ab \in R$.
(M1) (**Associative law**): $a(bc) = (ab)c$ for all $a, b, c \in R$.
Mixed axiom
(D) (**Distributive laws**): $(a + b)c = ac + bc$ and $c(a + b) = ca + cb$ for all $a, b, c \in R$.

The two distributive laws are sometimes called the **left** and **right distributive laws** respectively. Please do not try to remember which is which; I will not use these names.

The closure laws (A0) and (M0) are not strictly necessary: when we say that addition and multiplication are operations on R, it follows that the closure laws must hold! We will see the reason for requiring them when we come to look at subrings in Section 2.4.

Just as, when buying a personal computer, you are offered various extra features (more RAM, larger hard disk, etc.), so it is possible to have extra features in your ring if you are prepared to spend a bit more. Some of these are as follows. A ring R is a **ring with identity** if it satisfies

(M2) (**Identity law**): There exists $1 \in R$ with $1 \neq 0$ such that $a1 = 1a = a$ for all $a \in R$.

R is a **division ring** if it satisfies (M2) and also

(M3) (**Inverse law**): For all $a \in R$ with $a \neq 0$, there exists $b \in R$ with $ab = ba = 1$.

R is a **commutative ring** if it satisfies

(M4) (**Commutative law**): $ab = ba$ for all $a \in R$.

Finally, a commutative division ring is called a **field**.

Note that the extra multiplicative axioms are almost exact parallels of the additive axioms that we require in any ring. The exception is that, in the inverse law, we only require that non-zero elements have multiplicative inverses. We will see the reason for this soon.

Many authors make the convention that a ring must have an identity. In other words, they assume axiom (M2) along with (A0)–(A4), (M0), (M1), and (D).

In fact, some go so far as to use the word **rng** (sic) for a structure which I called a ring (that is, satisfying (A0)–(A4), (M0), (M1), and (D), so that they can use the term r**I**ng for 'rng with **I**dentity'.

Of course it is just convention. When different groups of mathematicians use different conventions, you are free to choose the one you like best, but you must accept that other people will do things differently. When we have learned a bit more about rings, I will explain why I took the decision I did on page 79.

The commutative law for addition follows from the other axioms for a ring with identity. The simple argument for this is outlined on page 69. A different proof is outlined in the solution to Exercise 2.7.

Remark Remember that the qualifying expressions in the terms 'commutative ring' and 'ring with identity' refer to the *multiplication*. The addition in a ring is always commutative, and there is always an identity (or zero) element for addition.

2.2 Examples of rings.

Example 1 Our prototype of a ring is the ring $\mathbb{Z}$ of integers. It is indeed a ring; in fact, it is a commutative ring with identity (but not a field, since, for example, there is no integer x such that $2x = 1$). I assume that all of these properties of integers are familiar to you. To give formal proofs, it is necessary to have a careful definition of the integers. This is done in courses on the Foundations of Mathematics: we will look at the arguments in Chapter 6.

Example 2 Other familiar number systems, such as $\mathbb{Q}$ (the rational numbers), $\mathbb{R}$ (the real numbers), and $\mathbb{C}$ (the complex numbers), are fields.

Example 3: Matrix rings Let R be any ring. Let $M_n(R)$ denote the set of all $n \times n$ matrices with elements in R. We can define addition and multiplication on $M_n(R)$ by rules which look exactly the same as those for matrices of real numbers, which we saw in the last chapter. That is, if $A = (a_{ij})$ and $B = (b_{ij})$ (this means that the element in row i and column j of A is a_{ij}, etc.), then

$$A + B = C = (c_{ij}), \text{ where } c_{ij} = a_{ij} + b_{ij},$$

$$A \cdot B = D = (d_{ij}), \text{ where } d_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

(Note that the rule for matrix addition depends on addition of ring elements, while the rule for matrix multiplication involves calculating n products and adding them up. There is a potential problem here, since we can only add two elements at a time: we will see in the next section that it does not matter how we perform the additions.)

It can be shown that $M_n(R)$ is a ring. (See Exercise 2.2 for the case $n = 2$.)

If R has an identity, then so does $M_n(R)$ (the usual identity matrix with 1 on the diagonal and 0 everywhere else). But $M_n(R)$ is not commutative except in trivial cases, and is never a division ring for $n > 1$.

Example 4: Polynomial rings For any ring R, the set $R[x]$ of all polynomials with coefficients in R is a ring. This is a generalisation of the familiar case of real polynomials. We will discuss exactly what a polynomial is, and how addition and multiplication should be defined in general, in Section 2.9.

Example 5: Finite rings A finite ring can be specified by giving operation tables for its addition and multiplication. For obvious reasons, these tables are usually called **addition tables** and **multiplication tables**.

For example, it can be shown that the structure given by

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

is a field. This can be proved directly, but it takes some work. For example, to verify the associative law (A1) from the tables, we have to substitute all possible values of a, b, and c. There are two possibilities for each of these, so $2^3 = 8$ instances of the law to be checked. Also, of course, eight instances of the associative law for multiplication, four of the commutative law, 16 of the distributive law .... For larger finite rings the situation is even worse. The moral is that, if all else fails, this method can be used; but usually it is better to have a more theoretical proof!

You probably recognised that the ring with the above tables is $\mathbb{Z}_2$, the integers mod 2. In fact, $\mathbb{Z}_m$ is a ring, for any positive integer m. But not all finite rings are of this kind. Here is one which is not.

$+$	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

$\cdot$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	a	0
b	0	b	0	b

Example 6: Zero rings Let R be a set with one binary operation $+$ satisfying axioms (A0)–(A4). (Later, we will see that such a thing is called an **abelian group**.) Is it possible to define a multiplication on R so that it becomes a ring? The answer is yes: it is always possible, by the trivial rule

$$ab = 0 \text{ for all } a, b \in R,$$

where 0 is the zero element given by (A2).

To prove this, we check the remaining axioms:

(M0): For all $a, b \in R$, $ab = 0 \in R$.

(M1): For all $a, b, c \in R$, $(ab)c = 0 = a(bc)$.

(D): For all $a, b, c \in R$, we have $(a+b)c = 0$, while $ac + bc = 0 + 0 = 0$, using property (A2). Similarly the other way round.

A ring constructed in this manner (one in which all products are zero) is called a **zero ring**. Such rings always exist, but they are not very exciting.

Example 7 The set of all even integers is a ring. It is commutative, but does not have an identity. (There is no even integer e such that $ex = x$ for all even integers x.)

Example 8: Boolean rings Just as Descartes aimed to turn geometry into algebra by setting up coordinates in the Euclidean plane, so Boole attempted to turn set theory (and logic) into algebra, as we see below. The main legacy of his attempt is that his name is familiar to every computer scientist.

Let X be a set, and let R denote the **power set** of X, the set of all subsets of X. (This is sometimes denoted by $\mathcal{P}(X)$.) We define operations on R as follows. For $A, B \subseteq X$, we let $A + B$ be the **symmetric difference** of A and B, the set of all elements lying in either A or B but not both. (This is sometimes written $A \triangle B$.) Also, we let $A \cdot B$ be the intersection $A \cap B$.

Now R is a ring. Let us check the axioms.

(A0): Clear.

(A1): Use a Venn diagram or truth table to show that $(A + B) + C$ and $A + (B + C)$ are both equal to the set of elements which are either in all three of the sets A, B, C or in exactly one of them.

(A2): $A + \emptyset = A$, since nothing is in the empty set; so $\emptyset$ is the zero element.

(A3): $A + A = \emptyset$, since there is no element which lies in A but not in both A and A.(!) So the inverse of A is A.

(A4): Clear.
(M0): Clear.
(M1): $(AB)C$ and $A(BC)$ both consist of the elements lying in all three sets.
(D): Prove this by means of a Venn diagram or truth table.

Now R is a commutative ring. It has an identity, namely the whole set X (since $X \cap A = A$ for any $A \subseteq X$). But it is not a division ring if X has more than one element: the equation $A \cap B = X$ can never hold if A is a proper subset of X.

A ring of this form is called a **Boolean ring.** For example, when $X = \{0, 1\}$, the addition and multiplication tables are as follows:

$+$	$\emptyset$	$\{0\}$	$\{1\}$	$\{0,1\}$
$\emptyset$	$\emptyset$	$\{0\}$	$\{1\}$	$\{0,1\}$
$\{0\}$	$\{0\}$	$\emptyset$	$\{0,1\}$	$\{1\}$
$\{1\}$	$\{1\}$	$\{0,1\}$	$\emptyset$	$\{0\}$
$\{0,1\}$	$\{0,1\}$	$\{1\}$	$\{0\}$	$\emptyset$

$\cdot$	$\emptyset$	$\{0\}$	$\{1\}$	$\{0,1\}$
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
$\{0\}$	$\emptyset$	$\{0\}$	$\emptyset$	$\{0\}$
$\{1\}$	$\emptyset$	$\emptyset$	$\{1\}$	$\{1\}$
$\{0,1\}$	$\emptyset$	$\{0\}$	$\{1\}$	$\{0,1\}$

2.3 Properties of rings. In this section we prove a few basic properties which follow from the ring axioms.

1. In a ring, we can only add elements two at a time. What if we want to add more than two elements? We have to put in brackets to convert the sum into a succession of pairwise additions. However, because of the associative law, the answer is the same no matter how we put in the brackets. For example, consider $a+b+c+d$. There are five possible ways of evaluating this, corresponding to the five bracketings $((a+b)+c)+d$, $(a+(b+c))+d$, $(a+b)+(c+d)$, $a+((b+c)+d)$, and $a+(b+(c+d))$. Now $(a+b)+c = a+(b+c)$, so the first two are equal. Similarly, $(b+c)+d = b+(c+d)$, so the fourth and fifth are equal. Now consider $(a+b)+(c+d)$. Putting $a+b = x$, this is $x+(c+d) = (x+c)+d = ((a+b)+c)+d$; similarly, putting $c+d = y$, it works out to $a+(b+(c+d))$. So all the expressions are equal.

We usually write $a+b+c+d$, leaving out the brackets.

In fact, the sum of any number of elements does not depend on the bracketing used to work it out. This might seem obvious to you as an extension of the above argument. But we can (and should) give a correct formal proof.

Proposition 2.1 *In a ring, the sum $a_1 + \cdots + a_n$ of any number of elements is independent of the bracketing used to work it out.*

Proof The proof is by induction on n. For $n = 1$ and $n = 2$, there is nothing to prove. For $n = 3$, there are just two possible bracketings, namely $(a_1 + a_2) + a_3$ and $a_1 + (a_2 + a_3)$, and the associative law tells us that they are equal. So let us assume that the result holds for sums of fewer than n terms, and prove it for sums of n terms. The induction hypothesis allows us to write $a_1 + \cdots + a_m$ for the sum of m terms whenever $m < n$.

Now consider two bracketings of the sum of $a_1, \ldots, a_n$. In the evaluation of each bracketing, at the last-but-one stage, we will have two expressions, the sum of $a_1, \ldots, a_k$, and the sum of $a_{k+1}, \ldots, a_n$, which are then added at the last stage. By the induction hypotheses, each of these smaller sums is independent of the bracketing, and so we can write the whole expression as

$$(a_1 + \cdots + a_k) + (a_{k+1} + \cdots + a_n),$$

where $0 < k < n$. Similarly, the other expression reduces to

$$(a_1 + \cdots + a_l) + (a_{l+1} + \cdots + a_n),$$

where $0 < l < n$.

If $k = l$, these expressions are clearly equal. So suppose not. We may assume that $k < l$. Now, again using the inductive hypothesis, we can write the first as

$$(a_1 + \cdots + a_k) + ((a_{k+1} + \cdots + a_l) + (a_{l+1} + \cdots + a_n)),$$

and the second as

$$((a_1 + \cdots + a_k) + (a_{k+1} + \cdots + a_l)) + (a_{l+1} + \cdots + a_n).$$

But these have the form $x + (y + z)$ and $(x + y) + z$, where

$$\begin{aligned} x &= a_1 + \cdots + a_k, \\ y &= a_{k+1} + \cdots + a_l, \\ z &= a_{l+1} + \cdots + a_n; \end{aligned}$$

by the associative law, they are equal. □

The argument does not depend on the fact that the operation is called 'addition', but only on the associative law. So the same is true, for example, for the operation of multiplication in a ring.

2. Axiom (A2) guarantees that a zero element exists. Could there be more than one? Suppose that z_1 and z_2 are two zero elements in a ring R; that is, for all $a \in R$,

$$a + z_1 = z_1 + a = a = a + z_2 = z_2 + a.$$

Then we have

$$z_1 = z_1 + z_2 = z_2.$$

So *the zero element is unique.* A very similar argument shows that, in a ring with identity, *the identity element is unique.*

3. It is also true that inverses (as given by (A3)) are unique. For suppose that b and c are both inverses of a. This means that

$$a + b = b + a = 0 = a + c = c + a.$$

Now

$$b + (a + c) = b + 0 = b,$$
$$(b + a) + c = 0 + c = c.$$

By the associative law, the left-hand sides are equal; so $b = c$.

We write the inverse of a as $-a$. Also, we abbreviate $a + (-b)$ to $a - b$. Notice that we are falling into mathematical bad habits here: the symbol $-$ is being used in the first place as a unary operator (taking a to $-a$), and in the second as a binary operator (combining a and b to form $a - b$). In fact, people are used to this double use and have no trouble with it, but many calculators have different buttons for the two different uses of $-$.

Similarly, in a division ring, the multiplicative inverse of a non-zero element a is unique (and is written a^{-1}, so that $aa^{-1} = a^{-1}a = 1$).

4. The **cancellation law** holds for addition:

(C) (**Cancellation Law**): If $a + c = b + c$, then $a = b$.

Proof Suppose that $a + c = b + c$. Add $-c$ to each side:

$$(a + c) - c = (b + c) - c,$$
$$a + (c - c) = b + (c - c),$$
$$a + 0 = b + 0,$$
$$a = b.$$

□

5. Here is the proof that the commutative law for addition follows from the other axioms in a ring with identity.

Expand $(1 + 1)(a + b)$ in two ways. We get

$$\begin{aligned}(1 + 1)(a + b) &= (1 + 1)a + (1 + 1)b \\ &= a + a + b + b,\end{aligned}$$

and

$$\begin{aligned}(1 + 1)(a + b) &= 1(a + b) + 1(a + b) \\ &= a + b + a + b.\end{aligned}$$

So $a + a + b + b = a + b + a + b$. Cancelling a from the front and b from the end gives $a + b = b + a$, as required.

6. For any $a \in R$, $a0 = 0a = 0$.

Proof

$$a(0 + 0) = a0 + a0.$$

Also, $0 + 0 = 0$, so

$$a(0 + 0) = a0 = a0 + 0.$$

Now use (C) to cancel $a0$ from the equation to give $a0 = 0$.

The proof that $0a = 0$ is similar. □

Remark This is the reason why axiom (M3) for a division ring only requires *non-zero* elements to have multiplicative inverses; for $0 \neq 1$, and $x0 = 0$ for all $X \in R$, so there cannot exist b with $b0 = 1$.

7. For any $a \in R$, $-(-a) = a$.

Proof $-(-a)$ is the inverse of $-a$; that is, it is an element which, when added to $-a$, gives zero. But we know that a added to $-a$ gives zero. Since inverses are unique, these two inverses of $-a$ must be equal. □

8. For any $a, b \in R$, $-(a + b) = -b - a$.

Proof We have to show that $-b - a$ is the inverse of $a + b$. So add it to $a + b$:

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

as required. □

2.4 Subrings. Let R be a ring. A **subring** of R is a subset $S \subseteq R$ which itself forms a ring (using the same operations as those in R).

Let us see what checking the axioms involves in this case.

(A0): We require closure, that is, for all $a, b \in S$, $a + b \in S$.

(A1): The associative law automatically holds for all $a, b, c \in S$, since it holds for all a, b, c in the larger set R.

(A2): We require that the zero element of R lies in S.

(A3): For each $a \in S$, we require that $-a \in S$.

(A4): This holds automatically, by the same argument as for (A1).

(M0): We require that S is closed under multiplication.

(M1): This is automatic, as for (A1). The same is true of (D).

We conclude that, of the eight axioms, four are automatically true, just because we are looking at a subset of a ring. (These are the axioms asserting that all elements satisfy some equation.) So we only have to require the two closure axioms, the zero and inverse axioms.

In fact, we can whittle these down to three:

Theorem 2.2 (First Subring Test) *A non-empty subset S of a ring R is a subring provided that, for all $a, b \in S$, we have $a + b, ab, -a \in S$.*

Proof We are given the closure and inverse axioms; we have to show that $0 \in S$. But S is non-empty, so take any element $a \in S$. Then, by assumption, $-a \in S$, and so $0 = a + (-a) \in S$, as required. □

We can do even better, reducing the number of tests to two: closure under subtraction and multiplication.

Theorem 2.3 (Second Subring Test) *A non-empty subset S of a ring R is a subring provided that, for all $a, b \in S$, we have $a - b, ab \in S$.*

Proof Suppose that S is closed under subtraction and multiplication. To show it is a subring, we verify the conditions of the First Subring Test. Take an element $a \in S$. Then $a - a = 0 \in S$; so $0 - a = -a \in S$, and the inverse law holds. Now take $a, b \in S$. Then $-b \in S$, and so $a - (-b) = a + b \in S$, and we have closure under addition. Closure under multiplication is given; so S is a subring. □

Example We find all the subrings of the ring $\mathbb{Z}$ of integers.

First, we show that, for any integer m, the set $m\mathbb{Z} = \{mx : x \in \mathbb{Z}\}$ of all multiples of m is a subring. Take $a, b \in m\mathbb{Z}$; let $a = mx$, $b = my$, for some integers x, y. Then

$$a - b = m(x - y) \in m\mathbb{Z},$$
$$ab = m(mxy) \in m\mathbb{Z},$$

so $m\mathbb{Z}$ passes the Second Subring Test.

Now we show that every subring of $\mathbb{Z}$ is of this form. So let S be a subring. Certainly $0 \in S$. If $S = \{0\}$, then $S = 0\mathbb{Z}$ is of the required form. So suppose not. If $n \in S$, then also $-n \in S$; so S must contain some positive integer. Let m be the smallest positive integer in S. We will prove that $S = m\mathbb{Z}$. Proving this equality involves showing that each element of one set is in the other and vice versa.

First, take any element of $m\mathbb{Z}$, say mx. If $x = 0$, then $mx = 0 \in S$. If $x > 0$, then $mx = m + m + \cdots + m$ (x terms), and $m \in S$; so $mx \in S$ by closure. If $x < 0$, let $x = -y$. Then $my \in S$ as above, and then $mx = -my \in S$.

Conversely, take any element of S, say n. By the Division Algorithm for integers, we can divide n by m, obtaining a quotient q and remainder r; thus, $n = mq + r$, and $0 \leq r < m$. Now $n \in S$ and $mq \in S$, so $r = n - mq \in S$. If $r > 0$, we have a contradiction to the fact that m is the smallest positive integer in S. So, necessarily, $r = 0$ and $n = mq \in m\mathbb{Z}$.

Remark If you are asked to prove that something is a ring, it is usually much easier to recognise that it is a subset of a structure known to be a ring, and then apply one of the subring tests, than it is to check the eight ring axioms directly. Bear this in mind when you tackle Problems 2.1 and 2.3.

Exercise 2.1 Which of the following sets are rings (with the usual addition and multiplication):

(a) the natural numbers;
(b) the real polynomials of degree at most n;
(c) all polynomials with integer coefficients;
(d) all polynomials with integer coefficients and constant term zero;
(e) all polynomials with integer coefficients and degree at most four;
(f) all real polynomials f such that $f(2) = 0$;
(g) all integers divisible by 3;
(h) all non-singular 2×2 real matrices;
(i) all complex numbers of the form $a + b\mathrm{i}$ for $a, b \in \mathbb{Z}$;
(j) all real functions of the form $f(x) = ax + b$ for $a, b \in \mathbb{R}$.

Exercise 2.2 Let R be a ring, and let $M_2(R)$ denote the set of all 2×2 matrices with elements from R. Define addition and multiplication of 2×2 matrices by the usual rules:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix},$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}.$$

Prove carefully that $M_2(R)$ is a ring.

Exercise 2.3 Which of the following sets of 2×2 matrices over the real numbers are rings (with the addition and multiplication defined in Problem 2.2)? Which are commutative? Which have an identity? Which are division rings?

(a) The set of all symmetric matrices (matrices A satisfying $A^\top = A$).
(b) The set of all skew-symmetric matrices (matrices A satisfying $A^\top = -A$).
(c) The set of all upper-triangular matrices (matrices of the form $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$).
(d) The set of all strictly upper-triangular matrices (matrices of the form $\begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}$).
(e) The set of matrices of the form $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$.

[The **transpose** $A^\top$ of a matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given by $A^\top = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$.]

Exercise 2.4 Prove that, in any ring R,

$$\begin{aligned}(a_1 + a_2 + \cdots + a_m)(b_1 + b_2 + \cdots + b_n) = {} & a_1b_1 + a_1b_2 + \cdots + a_1b_n \\ & + a_2b_1 + \cdots \\ & + a_mb_1 + \cdots + a_mb_n.\end{aligned}$$

Exercise 2.5 Let R be a ring. For a positive integer n, let $n \cdot x$ denote $x + \cdots + x$ (n terms). Prove that

(a) $(m+n)\cdot x = m\cdot x + n\cdot x$ and $(mn)\cdot x = m\cdot(n\cdot x)$.
(b) if 1 is an identity and $n\cdot 1 = 0$, then $n\cdot x = 0$ for all x.

Exercise 2.6 Let x and y be elements of a ring R, and suppose that $xy = yx$. Prove the **Binomial Theorem** for $n > 0$:

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} \cdot x^{n-i} y^i.$$

[As in the preceding question, $m\cdot x = x + \cdots + x$ (m terms).]

Exercise 2.7 Let R be a ring with identity element 1.

(a) Prove that $(-1)\cdot x = -x$, where -1 and $-x$ are the additive inverses of 1 and x.
(b) Show that $-(x+y) = -y - x$.
(c) Hence show that the commutativity of addition can be deduced from the other axioms for a ring with identity.

Show that this is false if no identity element exists. [*Hint*: Let all products be zero.]

Exercise 2.8 Let R be a ring in which every element x satisfies $x^2 = x$ (where x^2 means xx).

(a) By evaluating $(x+x)^2$, show that $x + x = 0$ for all $x \in R$.
(b) By evaluating $(x+y)^2$, show that R is commutative.

Remark Any Boolean ring satisfies the condition $x^2 = x$ for all $x \in R$. It can be shown that any finite ring satisfying this condition is a Boolean ring.

In the spirit of abstract algebra, we will re-define the term **Boolean ring** to mean a ring R with identity satisfying $x^2 = x$ for all $x \in R$.

Exercise 2.9 Let R and S be rings. Define operations on $R \times S$ (the set of ordered pairs) by the rules

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2),$$
$$(r_1, s_1)(r_2, s_2) = (r_1 r_2, s_1 s_2).$$

Prove that $R \times S$ is a ring. Show further that $R \times S$ is commutative if and only if R and S are commutative, and that $R \times S$ has an identity if and only if R and S do. Can $R \times S$ ever be a field?

Remark $R \times S$ is known as the **direct product** or **direct sum** of the two rings R and S.

Exercise 2.10 (a) Let R be a ring in which the elements are the integers, and the addition is the same as in $\mathbb{Z}$. Is it possible that the multiplication is different from that in $\mathbb{Z}$? Can you describe all such rings?

(b) (**) Let R be a ring in which the elements are the integers, and the multiplication is the same as in $\mathbb{Z}$. Is it possible that the addition is different from that in $\mathbb{Z}$?

(Part (b) is much more difficult than part (a); you are not expected to solve it at this stage. We will return to this later.)

Exercise 2.11 (∗) The field $\mathbb{C}$ of **complex numbers** was described in Chapter 1 as the set of all objects $a + b\mathrm{i}$, where a and b are real numbers, where addition and multiplication are defined according to 'the usual rules' subject to the extra condition that $\mathrm{i}^2 = -1$. Hamilton constructed a larger number system, the **quaternions**, as follows. The elements of $\mathbb{H}$ are all objects of the form $a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$. Addition and multiplication are defined according to the usual rules, subject to the condition that multiplication of the elements $1, \mathrm{i}, \mathrm{j}, \mathrm{k}$ works according to the following table:

$\cdot$	1	i	j	k
1	1	i	j	k
i	i	-1	k	$-\mathrm{j}$
j	j	$-\mathrm{k}$	-1	i
k	k	j	$-\mathrm{i}$	-1

Prove that $\mathbb{H}$ is a non-commutative division ring. [*Hint*: Define the **conjugate** of the element $z = a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$ to be $\overline{z} = a - b\mathrm{i} - c\mathrm{j} - d\mathrm{k}$.] Prove that

$$z\overline{z} = (a^2 + b^2 + c^2 + d^2) \cdot 1.$$

Homomorphisms and ideals

Two rings are essentially the same for the purposes of algebra (the technical term is 'isomorphic') if we can match up their elements in such a way that addition and multiplication correspond; that is, if a corresponds to a' and b to b', then $a + b$ corresponds to $a' + b'$ and ab to $a'b'$. In this section, we define a weaker relationship between rings, which merely asserts that they are somewhat alike.

2.5 Cosets. The first topic seems to be a digression, but its relevance will become clear soon. Let S be a subring of the ring R. We will partition R into subsets called **cosets** of S.

Define a relation E on the set R by the rule that $(a, b) \in E$ if $b - a \in S$. I claim that E is an equivalence relation. It is

- *reflexive*, since $a - a = 0 \in S$;
- *symmetric*, since if $b - a \in S$ then $a - b = -(b - a) \in S$;
- *transitive*, since if $b - a \in S$ and $c - b \in S$ then $c - a = (c - b) + (b - a) \in S$.

So it is indeed an equivalence relation.

By the Equivalence Relation Theorem, R is partitioned into equivalence classes $E(a)$, where $E(a) = \{b : (a, b) \in E\}$. These equivalence classes are called the **cosets** of S in R. We examine them a bit more closely, and observe that

$$E(a) = S + a = \{s + a : s \in S\}.$$

To see this, first take $b \in S + a$. Then $b = s + a$ for some $s \in S$, so $b - a = s \in S$, whence $(a, b) \in E$ and $b \in E(a)$. Conversely, if $b \in E(a)$, then $b - a \in S$. Putting $s = b - a$, we have $b = s + a \in S + a$ as required.

Example Let $R = \mathbb{Z}$ and $S = 4\mathbb{Z}$. Then

$$\begin{aligned} S = S + 0 &= \{\ldots, -8, -4, 0, 4, 8, \ldots\}, \\ S + 1 &= \{\ldots, -7, -3, 1, 5, 9, \ldots\}, \\ S + 2 &= \{\ldots, -6, -2, 2, 6, 10, \ldots\}, \\ S + 3 &= \{\ldots, -5, -1, 3, 7, 11, \ldots\}, \end{aligned}$$

and $S + 4 = S$, at which point the sequence repeats. (Compare the example in Section 1.4.3.)

More generally, if $R = \mathbb{Z}$ and $S = n\mathbb{Z}$, where n is a positive integer, then the coset $n\mathbb{Z} + a$ is the set of all integers congruent to a mod n. Thus, the cosets are the congruence classes mod n, and there are n of them altogether.

The element a is called a **coset representative** for the coset $S + a$. Note that the system is perfectly democratic: any element of a coset can serve as its representative. (Actually, this is very slightly misleading in one case. The subring S is a coset of itself, namely $S + 0$; and while we could use any of its elements as a representative, it is most natural to use the element 0.)

2.6 Homomorphisms and ideals. I introduce these two slightly strange words by means of an example. Suppose that I am short-sighted (actually this is correct), and that when I look at an integer I can only see whether it is even or odd. I will not know much about the integers, but I will know enough to make some consistent statements about addition and multiplication: for example, even plus even equals even. My knowledge can be summarised in tables:

$+$	even	odd
even	even	odd
odd	odd	even

$\cdot$	even	odd
even	even	even
odd	even	odd

.

This looks very much like the two-element ring of Example 5 in Section 2.2. The point is that it is indeed a ring, and captures a little bit of the 'shape' of the ring of integers.

We define a **homomorphism** from a ring R to a ring S to be a function or map $\theta : R \to S$ which satisfies

$$\begin{aligned} \theta(a + b) &= \theta(a) + \theta(b), \\ \theta(ab) &= \theta(a)\theta(b) \end{aligned}$$

for all $a, b \in R$. Note that the addition and multiplication on the left of these equations are the operations in R, while those on the right are the operations in S.

Having said this, I will perversely change notation right away. There are good reasons for writing the result of applying θ to a, not as $\theta(a)$, but as $a\theta$. With this notation, we say that the map θ is written **on the right**. One reason is that, in algebra, we often have to apply a function θ followed by another function ϕ: this is written as $a\theta\phi$, whereas if we wrote functions on the left we would have to say $\phi(\theta(a))$, and always remember to reverse the order. From now on, functions with an algebraic significance, such as homomorphisms, will always be written on the right; while functions with no such significance, such as polynomials, will be written on the left. So $f(x) = x^2 + 1$ is a polynomial, and $f\theta$ is the result of applying to it some homomorphism from the ring of all polynomials to another ring. Confused? Remember that not everybody uses this convention!

Let us rewrite the definition of a homomorphism in the new notation $\theta : R \to S$ is a homomorphism if

$$(a+b)\theta = a\theta + b\theta,$$
$$(ab)\theta = (a\theta)(b\theta).$$

The word *homomorphism* means 'similar shape'; this is meant to suggest that some of the 'shape' of the ring R is captured by S, as in our example. In this terminology, if S is the ring $\{0, 1\}$ of Example 5 of Section 2.2, then the function $\theta : \mathbb{Z} \to S$ defined by

$$n\theta = \begin{cases} 0 & \text{if } n \text{ is even} \\ 1 & \text{if } n \text{ is odd,} \end{cases}$$

is a homomorphism.

A homomorphism which is also a bijection (a one-to-one and onto function) is called an **isomorphism**. If there is an isomorphism from R to S, we say that the rings R and S are **isomorphic**. This means that they are 'matched up' by the function θ in such a way that the addition and multiplication are the same. So, from the point of view of abstract algebra, we will regard the two rings as being the same, even if their actual elements are quite different (one ring might consist of matrices and the other of polynomials, say). We denote 'R and S are isomorphic' by $R \cong S$.

Any homomorphism $\theta : R \to S$ has the additional properties

$$0\theta = 0,$$
$$(a-b)\theta = a\theta - b\theta$$

for all $a, b \in R$. The first equation follows from

$$0\theta + 0\theta = (0+0)\theta = 0\theta = 0\theta + 0$$

by using the Cancellation Law in S. The second from the fact that

$$(a-b)\theta + b\theta = (a-b+b)\theta = a\theta.$$

Our main task in this section is to see just how much a homomorphism blurs the structure of a ring, and how much shape is preserved.

Example We will find all homomorphisms from the ring $\mathbb{Z}$ to itself. Let θ be a homomorphism. Suppose that $1\theta = n$. Then $2\theta = (1+1)\theta = n+n = 2n$, and similarly (by induction), $m\theta = mn$ for all positive integers m. Moreover, $0\theta = 0$, and for positive m we have $(-m)\theta = -m\theta = -mn$. So θ multiplies every integer by n.

So far, we have only used the additive property. Now we turn to multiplication, and observe that

$$n = 1\theta = 1\theta \cdot 1\theta = n \cdot n = n^2,$$

so that $n = 0$ or $n = 1$. So there are only two homomorphisms, namely

- θ_0: $x \mapsto 0$ for all x;
- θ_1: $x \mapsto x$ for all x.

In fact, these rules define 'trivial' homomorphisms on any ring R. So our favourite ring $\mathbb{Z}$ is somewhat poor in homomorphisms: the only homomorphisms from $\mathbb{Z}$ to itself are the trivial ones possessed by all rings. (Of course, as we saw, there are homomorphisms from $\mathbb{Z}$ to other rings.)

A homomorphism $\theta : R \to S$ is a function, and so has an image and a kernel in the sense of Section 1.18. As promised there, we simplify the definition of the kernel slightly.

Definition Let $\theta : R \to S$ be a homomorphism of rings. The **image** of θ is

$$\mathrm{Im}(\theta) = \{s \in S : s = r\theta \text{ for some } r \in R\},$$

and the **kernel** of θ is

$$\mathrm{Ker}(\theta) = \{r \in R : r\theta = 0\}.$$

Remark In Section 1.18, the kernel of a function was defined to be the equivalence relation in which two elements are equivalent if they have the same image. So $\mathrm{Ker}(\theta)$ is the equivalence class containing 0 of the relation $\mathrm{KER}(\theta)$.

Proposition 2.4 *Let $\theta : R \to S$ be a ring homomorphism.*

(a) $\mathrm{Im}(\theta)$ *is a subring of* S.
(b) $\mathrm{Ker}(\theta)$ *is a subring of* R *which has the additional property that, for any* $x \in \mathrm{Ker}(\theta)$ *and* $r \in R$, *we have* $rx, xr \in \mathrm{Ker}(\theta)$.
(c) Two elements of R *are mapped to the same element of* S *under* θ *if and only if they lie in the same coset of* $\mathrm{Ker}(\theta)$.

Part (c) of this result states that the equivalence classes of the equivalence relation $\mathrm{KER}(\theta)$ are precisely the cosets of the subring $\mathrm{Ker}(\theta)$. This is why we

use the simpler definition: we can obtain the entire kernel (in the first sense) from the subring $\mathrm{Ker}(\theta)$. This also shows an important property of cosets.

Proof (a) We apply the subring test. Take $s_1, s_2 \in \mathrm{Im}(\theta)$. Then $s_1 = r_1\theta$ and $s_2 = r_2\theta$, for some $r_1, r_2 \in R$. Then

$$s_1 - s_2 = r_1\theta - r_2\theta = (r_1 - r_2)\theta \in \mathrm{Im}(\theta),$$
$$s_1 s_2 = (r_1\theta)(r_2\theta) = (r_1 r_2)\theta \in \mathrm{Im}(\theta);$$

so $\mathrm{Im}(\theta)$ is a subring of S.

(b) Similarly, take $r_1, r_2 \in \mathrm{Ker}(\theta)$. Then $r_1\theta = r_2\theta = 0$; so

$$(r_1 - r_2)\theta = r_1\theta - r_2\theta = 0 - 0 = 0,$$
$$(r_1 r_2)\theta = (r_1\theta)(r_2\theta) = 0 \cdot 0 = 0;$$

so $r_1 - r_2, r_1 r_2 \in \mathrm{Ker}(\theta)$, and $\mathrm{Ker}(\theta)$ is a subring.

Now we check the extra condition. Suppose that $x \in \mathrm{Ker}(\theta)$ and $r \in R$. Then

$$(rx)\theta = (r\theta)(x\theta) = r\theta \cdot 0 = 0,$$

and so $rx \in \mathrm{Ker}(\theta)$. Similarly, $xr \in \mathrm{Ker}(\theta)$.

(c) Suppose that $r_1\theta = r_2\theta$. Then $(r_2 - r_1)\theta = 0$, so $x = r_2 - r_1 \in \mathrm{Ker}(\theta)$; then $r_2 \in \mathrm{Ker}(\theta) + r_1$, so r_1 and r_2 lie in the same coset of $\mathrm{Ker}(\theta)$. Conversely, if r_1 and r_2 lie in the same coset, say $r_2 = r_1 + x$ with $x \in \mathrm{Ker}(\theta)$; then $r_2\theta = r_1\theta + x\theta = r_1\theta$, since $x \in \mathrm{Ker}(\theta)$. □

The extra property of the subring $\mathrm{Ker}(\theta)$ is so important that it is given a special name. An **ideal** of a ring R is a subring S of R such that, for any $s \in S$ and $r \in R$, we have $rs, sr \in S$. The term was invented by Kummer, who invented 'ideal numbers' in an attempt to correct a mistake in his attempted proof of Fermat's Last Theorem. Unfortunately, he did not succeed, and we had to wait another hundred years until Fermat's Last Theorem was proved; but the concept of an ideal is crucial for ring theory.

To test for an ideal, we should test for a subring and then check the extra condition. But we can simplify things:

Theorem 2.5 (Ideal Test) *A non-empty subset S of a ring R is an ideal of R if and only if*

(a) for all $s_1, s_2 \in S$, we have $s_1 - s_2 \in S$;
(b) for all $s \in S$ and $r \in R$, we have $rs, sr \in S$.

Proof All that is missing is closure under multiplication; but this is just the special case of (b) corresponding to the case in which $r \in S$. □

So we can say more briefly:

The kernel of a homomorphism $\theta : R \to S$ is an ideal of R.

Example We found in Section 2.4 that the subrings of $\mathbb{Z}$ are all the sets of the form $n\mathbb{Z}$ for $n \in \mathbb{Z}$, where $n\mathbb{Z}$ consists of all multiples of n. Now all of these subrings are ideals. For take $r \in \mathbb{Z}$ and $s \in n\mathbb{Z}$, say $s = nx$ for some $x \in \mathbb{Z}$. Then $rs = sr = n(rx) \in n\mathbb{Z}$. So we have found all the ideals of $\mathbb{Z}$.

2.7 Should a ring have an identity? What are the reasons for not requiring the existence of an identity in a ring?

First, it is convenient that an ideal is a particular kind of subring. But if rings are required to have identities, then all their ideals (with one exception) fail to be subrings! (In the next chapter, we will see a close analogy between groups, subgroups, and normal subgroups on one hand, and rings, subrings, and ideals on the other. This analogy would fail if ideals were not subrings.)

Proposition 2.6 *Let R be a ring with identity element 1, and I an ideal containing 1. Then $I = R$.*

Proof For any element $r \in R$, we have $r = 1 \cdot r \in I$. □

Second, there are several important examples of rings which are used in various branches of mathematics and which do not contain identities. For example,

- $\mathcal{C}_0(\mathbb{R})$, the ring of continuous real-valued functions with bounded support;
- the direct sum of an infinite collection of rings (even if all the factors have identities).

Third, the argument that a more general definition is better. In topology, there was a debate over whether a topological space should be required to satisfy the 'Hausdorff condition' or not; this was resolved in favour of the more general approach.

Finally, the traditional defence: if a ring does not have an identity, we can put one in!

Proposition 2.7 *Let R be a ring. Then there is a ring R^* with identity which contains a subring isomorphic to R.*

Proof Let

$$R^* = R \times \mathbb{Z} = \{(r, n) : r \in R, n \in \mathbb{Z}\},$$

and define addition and multiplication on R^* by the rules

$$(r_1, n_1) + (r_2, n_2) = (r_1 + r_2, n_1 + n_2),$$
$$(r_1, n_1) \cdot (r_2, n_2) = (r_1 r_2 + n_2 r_1 + n_1 r_2, n_1 n_2),$$

where the product nr of an integer and a ring element is defined as in Exercise 2.5 for positive n (that is, the sum of n copies of r), with $0r = 0$ and $(-m)r = -(mr)$

for $n = -m < 0$. [*Warning*: This is not the direct product of rings defined in Exercise 2.9.]

Now a small amount of checking shows that

- R^* is a ring;
- $(0, 1)$ is the identity of R^*;
- $S = \{(r, 0) : r \in R\}$ is a subring of R^* (in fact, it is an ideal);
- the map $r \mapsto (r, 0)$ is an isomorphism from R to S.

Of course, if R already has an identity element, then this element is no longer the identity of R^*. □

2.8 Factor rings and isomorphism theorems. For any integer n, the set $n\mathbb{Z}$ of multiples of n is an ideal of $\mathbb{Z}$ (and all ideals are of this form). We saw earlier that the cosets of $n\mathbb{Z}$ are just the congruence classes mod n. But we can do more: we can add and multiply integers mod n. This means that, if we take two congruence classes $n\mathbb{Z} + x$ and $n\mathbb{Z} + y$, then the sum of any integer from the first class and any integer from the second will always lie in the congruence class $n\mathbb{Z} + (x + y)$; and, similarly, the product of integers from these classes will lie in $n\mathbb{Z} + xy$. In this way, we can define operations of addition and multiplication on the **set** of congruence classes mod n (a finite set with n elements). With these operations, the set of congruence classes becomes a ring.

This all works much more generally; for any ideal I of a ring R, it is possible to make the set of cosets of I in R into a ring, called the 'factor ring' of R by I and denoted R/I, as follows.

Definition Let I be an ideal in the ring R. The **factor ring** or **quotient ring** R/I is the set of cosets of I in R, with operations of addition and multiplication defined by

$$(I + x) + (I + y) = I + (x + y),$$
$$(I + x)(I + y) = I + xy.$$

Theorem 2.8 *The factor ring, as defined above, is indeed a ring.*

Proof Before we verify the axioms, there is one very important thing to check: that the definition is a good one. On the face of it, the definition depends on the choice of coset representatives. It is not clear that, if x_1 and x_2 are two representatives of a coset of I, and y_1 and y_2 are two representatives of another coset, then $x_1 + y_1$ and $x_2 + y_2$ lie in the same coset (and similarly for multiplication). Suppose that we have such elements. Then $x_2 = x_1 + a$ and $y_2 = y_1 + b$, for some $a, b \in I$. Then

$$(x_2 + y_2) - (x_1 + y_1) = (x_1 + a + y_1 + b) - (x_1 + y_1) = a + b \in I,$$
$$(x_2 y_2) - (x_1 y_1) = (x_1 + a)(y_1 + b) - (x_1 y_1) = x_1 b + a y_1 + ab \in I,$$

where, in the last step, $ab \in I$ since I is a subring, and $x_1 b, a y_1 \in I$ since I is an ideal. So the operations on R/I are indeed well defined.

Now the rest of the proof involves verifying the axioms, which is routine. The closure laws need no proof, since we have well-defined operations. For the associative law (A1), we have

$$\begin{aligned}((I+x)+(I+y))+(I+z) &= (I+(x+y))+(I+z)\\ &= I+((x+y)+z)\\ &= I+(x+(y+z))\\ &= (I+x)+(I+(y+z))\\ &= (I+x)+((I+y)+(I+z)).\end{aligned}$$

The proofs of (A4), (M1), and (D) are very similar. The zero element is $I+0=I$, and the inverse of $I+x$ is $I+(-x)$ (which, of course, we write as $I-x$). □

You were warned in the last chapter that a set of sets is difficult to think about, especially if we have to perform operations on its elements. Here is precisely that situation. But it is so important that it is worth taking some trouble to grasp the ideas.

You may recognise that, if R is the ring $\mathbb{Z}$ of integers and I is the set $m\mathbb{Z}$ of all multiples of the positive integer m, then the ring R/I is precisely the structure we called 'the integers mod m' in Chapter 1, and denoted by $\mathbb{Z}_m$. We will use the same notation here. The integers mod m provide a very good example of a factor ring.

The factor ring comes as the image of a natural homomorphism, as follows: Remember that the *elements* of R/I are the *cosets* of I in R. Now define a map $\theta : R \to R/I$ by the rule that $x\theta = I + x$ for all $x \in R$. Checking that θ is a homomorphism and is straightforward (we could say that the definitions of addition and multiplication in R/I were chosen to make this work):

$$\begin{aligned}(x+y)\theta = I+(x+y) = (I+x)+(I+y) = x\theta + y\theta,\\ (xy)\theta = I+xy = (I+x)(I+y) = (x\theta)(y\theta).\end{aligned}$$

The image of θ is R/I, since every coset has the form $I+x$ for some $x \in R$. What is the kernel of θ? Since the zero element of R/I is the coset I, we have

$$\mathrm{Ker}(\theta) = \{x \in R : I+x = I\} = \{x \in R : x \in I\} = I.$$

We call the map θ the **canonical homomorphism** from the ring R to its factor ring R/I. Hence we have proved the following:

Theorem 2.9 *The canonical homomorphism $\theta : R \to R/I$ defined by $x\theta = I+x$ for $x \in R$ is indeed a homomorphism; its image is R/I and its kernel is I.*

Armed with the concept of factor rings and the canonical homomorphism, we can return to our analysis of the image and kernel of an arbitrary homomorphism.

Theorem 2.10 (First Isomorphism Theorem) *Let $\theta : R \to S$ be a ring homomorphism. Then*

(a) $\mathrm{Im}(\theta)$ *is a subring of* S*;*
(b) $\mathrm{Ker}(\theta)$ *is an ideal of* R*;*
(c) $R/\mathrm{Ker}(\theta) \cong \mathrm{Im}(\theta)$*.*

Proof We have already shown (a) and (b). For (c), there is only one reasonable definition of a map ϕ from R/I to S, where $I = \mathrm{Ker}(\theta)$: we must put $(I + x)\phi = x\theta$ for all $x \in R$. As usual, we have to show that this is well defined. So let x_1 and x_2 be representatives of the same coset of I, so that $x_2 = x_1 + a$ for some $a \in I$. Then

$$x_2\theta = (x_1 + a)\theta = x_1\theta + a\theta = x_1\theta,$$

since $a\theta = 0$; so $(I + x_1)\phi = (I + x_2)\phi$, and ϕ is indeed well defined.

To show that ϕ is a homomorphism, we have

$$(I + (x + y))\phi = (x + y)\theta = x\theta + y\theta = (I + x)\phi + (I + y)\phi,$$
$$(I + xy)\phi = (xy)\theta = (x\theta)(y\theta) = ((I + x)\phi)((I + y)\phi).$$

Now ϕ is clearly onto $\mathrm{Im}(\theta)$, since for any $s \in \mathrm{Im}(\theta)$ we have $s = x\theta = (I+x)\phi$. Finally, suppose that $(I + x)\phi = (I + y)\phi$. Then $x\theta = y\theta$, so $(y - x)\theta = 0$; thus $y - x \in \mathrm{Ker}(\theta)$, and x and y represent the same coset of $\mathrm{Ker}(\theta)$. □

There are two further 'Isomorphism Theorems' relating a ring R to a factor ring R/I.

Theorem 2.11 (Second Isomorphism Theorem) *Let I be an ideal of R. There is a one-to-one correspondence between the set of subrings of R which contain I and the set of subrings of R/I. Under this correspondence, ideals of R containing I correspond to ideals of R/I.*

Proof If S is a subring of R containing I, then any coset of I with a representative in S is completely contained in S. (For, if $I \subseteq S$ and $x \in S$, then $I + x \subseteq S$ by closure of S.) Moreover, I is an ideal of S, since it is closed under subtraction and under multiplication by elements of S. So the factor ring S/I is the set of all cosets of I in S, and is a subring of R/I (as it is a ring in its own right). Conversely, let T be a subring of R/I. Then T is a set of cosets of I; the union of all these cosets is a subset $\hat{T}$ of R, which is easily seen to be a subring of R containing I. Hence we have the one-to-one correspondence. The further statement about ideals is an easy exercise. □

Theorem 2.12 (Third Isomorphism Theorem) *Let I be an ideal of R and S a subring of R. Then*

(a) $I + S = \{a + s : a \in I, s \in S\}$ *is a subring of R containing I;*
(b) $I \cap S$ *is an ideal of S;*
(c) $S/(I \cap S) \cong (I + S)/I$*.*

Proof All this can be proved directly; but a little trick, based on the natural homomorphism $\theta : R \to R/I$, makes it easier. We let ϕ be the restriction of θ to S. That is, ϕ maps S to R/I, and the value $s\phi$ for $s \in S$ is just $s\theta$. (We simply forget how θ acts on elements outside S.) Clearly, ϕ is a homomorphism: the two conditions in the definition hold for arbitrary elements of R, and so certainly for all elements of S.

(a) What is the image of ϕ? We see that $\mathrm{Im}(\phi)$ consists of all cosets $I + s$ for which the representative is in S. These form a subring of R/I, by Theorem 2.10(a). The union of all these cosets is the set

$$\{a + s : a \in I, s \in S\} = I + S$$

by Theorem 2.11; so $I + S$ is a subring of R which contains I. Incidentally, we see that $\mathrm{Im}(\phi) = (I + S)/I$.

(b) What is the kernel of ϕ? We see that $\mathrm{Ker}(\phi)$ consists of all the elements of S mapped to zero by θ. Since $\mathrm{Ker}(\theta) = I$, we have $\mathrm{Ker}(\phi) = I \cap S$, which is thus an ideal of S, by Theorem 2.10(b).

(c) By Theorem 2.10(c), $S/\mathrm{Ker}(\phi) \cong \mathrm{Im}(\phi)$; that is, $S/(I \cap S) \cong (I + S)/I$, as required. □

These theorems are quite abstract, and the proofs are very condensed. Here is an example in detail.

Example Let $R = \mathbb{Z}$, and let I be the ideal $4\mathbb{Z}$. The cosets of I are the congruence classes mod 4. For simplicity, we will write the class $4\mathbb{Z}+k$ as k (being careful to distinguish between the integer and the coset). Now the addition and multiplication tables of $\mathbb{Z}_4 = \mathbb{Z}/4\mathbb{Z}$ are as follows (ignore the underlines for the moment):

$+$	$\underline{0}$	1	$\underline{2}$	3
$\underline{0}$	$\underline{0}$	1	$\underline{2}$	3
1	1	2	3	0
$\underline{2}$	$\underline{2}$	3	$\underline{0}$	1
3	3	0	1	2

$\cdot$	$\underline{0}$	1	$\underline{2}$	3
$\underline{0}$	$\underline{0}$	0	$\underline{0}$	0
1	0	1	2	3
$\underline{2}$	$\underline{0}$	2	$\underline{0}$	2
3	0	3	2	1

Now $2\mathbb{Z}$ is a subring of $\mathbb{Z}$ containing $4\mathbb{Z}$: the corresponding subring of $\mathbb{Z}/4\mathbb{Z}$ is the set of cosets containing even numbers. These are the underlined cosets in the tables above: inspection of the tables shows that we do indeed have closure, so that $\{0, 2\}$ is a subring of $\mathbb{Z}/4\mathbb{Z}$, as it should be by Theorem 2.11. (Indeed, it is an ideal, also in accordance with that Theorem.)

Let $S = 6\mathbb{Z}$, a subring of $R = \mathbb{Z}$, and I the ideal $4\mathbb{Z}$. Then

$$I + S = \{4x + 6y : x, y \in \mathbb{Z}\} = 2\mathbb{Z},$$

the subring of R containing I described above. Also, $4\mathbb{Z} \cap 6\mathbb{Z} = 12\mathbb{Z}$, since an integer is divisible by both 4 and 6 if and only if it is divisible by 12. So Theorem 2.12 asserts that $6\mathbb{Z}/12\mathbb{Z} \cong 2\mathbb{Z}/4\mathbb{Z}$. The second of these factor rings

consists of the underlined elements in the above tables; the first has the following tables:

$+$	0	6
0	0	6
6	6	0

$\cdot$	0	6
0	0	0
6	0	0

(Note that $6 \cdot 6 = 36 \equiv_{12} 0$.) Inspection shows that the two factor rings are indeed isomorphic, by the correspondence $0 \leftrightarrow 0$, $2 \leftrightarrow 6$.

2.9 Polynomials. Like sets, polynomials are easy to understand, but difficult to define; we must make the attempt.

Usually, a polynomial is written as a sum of terms, where each term is a product of a coefficient and a power of an 'indeterminate' x. Traditionally, the coefficients are real numbers, and a polynomial is regarded as a function from $\mathbb{R}$ to $\mathbb{R}$. In keeping with the spirit of abstract algebra, we allow the elements of any ring R as coefficients; and we do not care what a polynomial really is, as we are only interested in the rules for adding and multiplying polynomials. (In fact, over some rings, different polynomials define the same function. We saw in Section 2.2 that $x^2 = x$ for all elements x of a Boolean ring R; so the polynomials x and x^2 would define the same function.)

Clearly, a polynomial is specified by giving its coefficients. But even these are not uniquely determined. If a polynomial has degree n, we can add to it an extra term $0x^{n+1}$ without changing it. Accordingly, we allow a polynomial to have infinitely many terms, but specify that in all but a finite number of them the coefficient is zero.

Now we are ready for the formal definition.

A **polynomial** over a ring R is an infinite sequence $(a_0, a_1, a_2, \ldots)$ of elements of R, indexed by the non-negative integers, with the property that there exists an integer n such that $a_i = 0$ for all $i > n$. In accordance with the usual notation, we write the sequence $(a_0, a_1, a_2, \ldots)$ as $a_0 + a_1 x + a_2 x^2 + \cdots$, or (if n is as in the definition) as $\sum_{i=0}^{n} a_i x^i$.

Addition and multiplication of polynomials are defined by the 'usual' rules (essentially the ones we saw in Chapter 1):

$$\left(\sum a_i x^i\right) + \left(\sum b_i x^i\right) = \sum c_i x^i, \text{ where } c_i = a_i + b_i,$$

$$\left(\sum a_i x^i\right) \cdot \left(\sum b_i x^i\right) = \sum d_i x^i, \text{ where } d_i = \sum_{j=0}^{i} a_j b_{i-j}.$$

We let $R[x]$ denote the set of polynomials over R, with the above addition and multiplication.

Theorem 2.13 *For any ring R, $R[x]$ is a ring. It is commutative if and only if R is commutative; it has an identity if and only if R has an identity; but it is never a division ring.*

The proof involves checking all the axioms; it will not be given here. The important thing to note is that the closure laws hold: adding and multiplying polynomials produce a sequence which has only finitely many non-zero terms, hence again a polynomial. Indeed, if we define the **degree** $\deg(f)$ of a non-zero polynomial to be the greatest integer n for which the coefficient of x^n is non-zero, then we have

$$\begin{aligned} \deg(f+g) &\leq \max(\deg(f), \deg(g)), \\ \deg(fg) &\leq \deg(f) + \deg(g). \end{aligned}$$

(We do not define the degree of the zero polynomial, since it has no non-zero terms at all. Some people would define its degree to be $-\infty$, while others would say -1; but these are mere conventions.)

A **constant polynomial** is a polynomial $\sum a_i x^i$ with $a_i = 0$ for $i > 0$. In other words, the constant polynomials are the zero polynomial and the polynomials whose degree is zero. They form a subring of $R[x]$ isomorphic to R. Often, we don't distinguish carefully between the ring element r and the constant polynomial $r = \sum a_i x^i$ with $a_0 = r$ and $a_i = 0$ for $i > 0$.

Remarks 1. If we consider all the infinite sequences of elements of R, without imposing the restriction that only finitely many are non-zero, and use the same definitions of addition and multiplication, we obtain another important ring, the **formal power series ring** over R, denoted $R[[x]]$. (The word 'formal' signifies that we do not attempt to sum the power series, and are not concerned with questions of convergence.)

2. Here is another definition of polynomials, which avoids the need to consider infinite sequences at the expense of another complication. Let X denote the set of all *finite* sequences $(a_0, a_1, \ldots, a_n)$ of elements of R. (The number n can take any value. In particular, we include the empty sequence, which has no terms at all.) Now we define a relation E on X by the rule that $(s,t) \in E$ if t can be obtained from s by either adding or deleting any number of zeros from the right-hand end of the sequence. It can be shown that E is an equivalence relation; its equivalence classes are **polynomials**. To add or multiply polynomials f and g, we choose representative sequences $s = (a_0, \ldots, a_n)$ and $t = (b_0, \ldots, b_m)$ from the equivalence classes f and g. We may assume that $m = n$, by adding zeros to the shorter sequence. Now define $f + g$ to be the equivalence class of $s + t$, and fg the equivalence class of st, where addition and multiplication of sequences is defined as before. It can be shown that these operations do not depend on the choice of representatives of the equivalence classes, so that they are well defined; and that, with these operations, the set of equivalence classes is a ring. Furthermore, this ring is isomorphic to the ring of infinite sequences which we defined before.

3. The upshot of this section is that you already understand polynomials, and you should think of them just as you did before; but they can be put on a proper theoretical basis, with some work.

Exercise 2.12 Let I be an ideal of a ring R. Prove that $M_n(I)$ (the set of $n \times n$ matrices with elements in I) is an ideal of $M_n(R)$. (For an easier question, do the case $n = 2$.) Prove also that $M_n(R)/M_n(I) \cong M_n(R/I)$.

Exercise 2.13 Let I be an ideal in a commutative ring R. Prove that $I[x]$ (the ring of polynomials over I) is an ideal in $R[x]$. Prove also that $R[x]/I[x] \cong (R/I)[x]$.

Exercise 2.14 $(*)$ Let R be a ring with identity. Suppose that J is an ideal in $M_n(R)$. Prove that there is an ideal I of R such that $J = M_n(I)$. [*Hint:* let E_{ij} be the **matrix unit** with 1 in row i and column j and 0 everywhere else. Prove that, if $A = (a_{ij})$, then $E_{ki}AE_{jl}$ has entry a_{ij} in the (k, l) position and zeros elsewhere. Now let I be the set of all elements of R which appear as an entry in some matrix of J. Show that, for any $r \in I$, the matrix with r in the top-left corner and 0 elsewhere belongs to J. Hence show that I is an ideal, and that $J = M_n(I)$.]

Exercise 2.15 (a) Show that, in the ring $\mathbb{Z}$, $m\mathbb{Z}$ contains $n\mathbb{Z}$ if and only if m divides n.

(b) How many ideals does the ring $\mathbb{Z}_{60}$ have? How many of these ideals are maximal (in the sense that I is maximal if $I \neq R$ but no ideal J satisfies $I \subset J \subset R$)?

(c) Repeat part (b) for the ring $\mathbb{Z}_n$, where $n = p_1^{a_1} \cdots p_r^{a_r}$ and $p_1, \ldots, p_r$ are distinct primes, and $a_1, \ldots, a_r$ are positive integers.

Exercise 2.16 (a) Prove that the **Gaussian integers**, the complex numbers of the form $a + b\mathrm{i}$, where a, b are integers, form a subring of $\mathbb{C}$.

(b) Prove that the **Eisenstein integers**, the complex numbers of the form $a + b\sqrt{-3}$, where either a, b are integers or $a - \frac{1}{2}, b - \frac{1}{2}$ are integers, form a subring of $\mathbb{C}$. (So, for example, $1 + \sqrt{-3}$ and $-\frac{1}{2} + \frac{5}{2}\sqrt{-3}$ are Eisenstein integers but $\frac{1}{2} - \sqrt{-3}$ is not.)

Exercise 2.17 Let R be a commutative ring and $u \in R$. Show that the map $\theta : R[x] \to R$ defined by 'substituting u for x'; that is, $\sum a_i x^i \mapsto \sum a_i u^i$, is a homomorphism.

Exercise 2.18 Let R be the ring $\mathbb{R}[x]$ of all real polynomials. Define a function $\theta : R \to \mathbb{C}$ by the rule that $f\theta = f(\mathrm{i})$. Prove that θ is a homomorphism, that its image is $\mathbb{C}$, and that its kernel is the ideal $(x^2 + 1)R$ consisting of all polynomials divisible by $x^2 + 1$. Hence show that

$$\mathbb{R}[x]/(x^2 + 1)\mathbb{R}[x] \cong \mathbb{C}.$$

Exercise 2.19 Construct a homomorphism from $\mathbb{Z}_{mn}$ to $\mathbb{Z}_n$, for any positive integers m, n.

Exercise 2.20 Let Y be a subset of the set X. Let $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ be the Boolean rings of subsets of X and Y, respectively. Show that the map $\theta : \mathcal{P}(X) \to \mathcal{P}(Y)$ defined by $A\theta = A \cap Y$ is a homomorphism, and find its image and kernel.

Exercise 2.21 Let R be the ring of real upper-triangular 2×2 matrices (those of the form $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$ for $a, b, c \in \mathbb{R}$). Let I be the set of strictly upper-triangular matrices (of the form $\begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix}$) and S the set of diagonal matrices (of the form $\begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}$).

(a) Prove that I is an ideal of R.
(b) Prove that S is a subring of R. Is it an ideal?
(c) Prove that R/I is isomorphic to S.

Factorisation

One of the most important properties of the integers is the so-called Fundamental Theorem of Arithmetic, which asserts that any integer can be factorised into primes in an essentially unique way. We want to examine the rings in which such a result could hold.

2.10 Zero-divisors and units. In contrast to the situation in $\mathbb{Z}$, it can happen in an arbitrary ring that the product of two non-zero elements is zero. For example, $2 \cdot 2 = 0$ in $\mathbb{Z}_4$, and $\left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right)\left(\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right) = \left(\begin{smallmatrix}0&0\\0&0\end{smallmatrix}\right)$ in $M_2(\mathbb{R})$. For much of the rest of this chapter, we are especially interested in rings in which this does not happen. Accordingly, we define it away:

- A **zero-divisor** in a ring R is a non-zero element $a \in R$ such that there exists a non-zero element $b \in R$ with $ab = 0$.
- An **integral domain** is a commutative ring with identity which has no zero-divisors.

Strictly speaking, in the first definition, a is a **left zero-divisor**, and b is a **right zero-divisor**; but, as the second definition suggests, we are mostly interested in commutative rings, and in these, the concepts of left and right zero-divisor coincide.

The condition 'no zero-divisors' can also be stated in the form: *if* $ab = 0$*, then either* $a = 0$ *or* $b = 0$. Thus, $\mathbb{Z}$, our prototype of a ring, is also our prototype of an integral domain, as the name would suggest. Integral domains have many nice properties. For example, there is a multiplicative version of the **cancellation law**:

(C′) In an integral domain, if $ab = ac$ and $a \neq 0$, then $b = c$.

For, if $ab = ac$, then $a(b - c) = 0$; in an integral domain, if $a \neq 0$, this implies that $b - c = 0$; that is, $b = c$.

Let R be a ring with identity. The element $a \in R$ is a **unit** if there exists $b \in R$ with $ab = ba = 1$. The element b is called the **inverse** of a, and is written a^{-1}. It is unique; for if b and c are both inverses of a, then

$$c = 1c = (ba)c = b(ac) = b1 = b.$$

You should compare this with the proof of uniqueness of additive inverses in Section 2.3.

Proposition 2.14 *Let R be a ring with identity.*

(a) The identity is a unit; it is equal to its inverse.
(b) If a is a unit, then so is a^{-1}; its inverse is a.
(c) If a and b are units, then so is ab; its inverse is $b^{-1}a^{-1}$.

Proof (a) $1 \cdot 1 = 1$.
(b) This is shown by the equations $aa^{-1} = a^{-1}a = 1$.
(c) We have

$$(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = a1a^{-1} = aa^{-1} = 1,$$

and, similarly, $(b^{-1}a^{-1})(ab) = 1$. □

Here is how Hermann Weyl explains part (c) of this proposition in his book *Symmetry* (1952).

> With this rule, although perhaps not with its mathematical expression, you are all familiar. When you dress, it is not immaterial in which order you perform the operations; and when in dressing you start with the shirt and end up with the coat, then in undressing you observe the opposite order; first take off the coat and the shirt comes last.

Examples 1. A ring with identity is a division ring if and only if every non-zero element is a unit.

2. The units in $\mathbb{Z}$ are 1 and -1.

3. In the next proposition, we find the zero-divisors and units in the ring $\mathbb{Z}_n$ of integers mod n, where $n > 1$.

Proposition 2.15 *An element $x \neq 0$ of $\mathbb{Z}_n$ is a zero-divisor if and only if x and n have greatest common divisor (g.c.d.) greater than 1; it is a unit if and only if x and n have greatest common divisor 1.*

In other words, in $\mathbb{Z}_n$, every non-zero element is either a zero-divisor or a unit (but not both, see Exercise 2.22).

Proof Suppose that $d = \gcd(x, n)$ is the greatest common divisor of x and n.

(a) If $d > 1$, then (n/d) is a non-zero element of $\mathbb{Z}_n$; and $x(n/d) = (x/d)n \equiv_n 0$.

(b) Suppose that $d = 1$. By the Euclidean algorithm, there are integers p and q such that $xp + nq = 1$. But this means that $xp = px \equiv_n 1$, so that x is a unit (and p is its inverse).

Conversely, if x is a zero-divisor, then x is not a unit, so d is not 1 — that is, $d > 1$ — and similarly for (b). □

Two elements a, b of the integral domain R are said to be **associates** if there is a unit $u \in R$ such that $b = au$. Note that, by the above Proposition, it follows that being associates is an equivalence relation: it is

- *reflexive*, since $a = a1$;
- *symmetric*, since $b = au$ implies $a = bu^{-1}$;
- *transitive*, since $b = au$ and $c = bv$ imply $c = a(uv)$.

(Here u and v are units; and the proposition shows that 1, u^{-1} and uv are units.)

By the Equivalence Relation Theorem, R is partitioned into equivalence classes, called **associate classes**.

For example, in $\mathbb{Z}$, the associate classes are the sets $\{n, -n\}$ for all non-negative integers n.

2.11 Irreducibles and factorisation. In this section, we examine the possibility of factorising elements of a ring into 'irreducible' elements (which cannot themselves be further factorised), and look at a special class of rings in which the analogue of the Fundamental Theorem of Arithmetic holds.

First, we will make some simplifying assumptions about the ring R. We always assume that R is commutative, so that we can regard ab and ba as essentially the same factorisation of a ring element. (So, in a factorisation, we do not care about the order of the factors.) Also, we exclude divisors of zero. For, if $ab = 0$, then $ac = a(b+c)$ for any element c, and there is little chance of unique factorisations.

Accordingly, we assume, in this section and the next two, that

R is an integral domain.

Also, units provide another problem. In $\mathbb{Z}$, we regard $2 \cdot 3$ and $(-2) \cdot (-3)$ as 'essentially the same' factorisation of 6. More generally, if $x = ab$, and u is a unit with inverse v, then $x = (au)(vb)$, and we want to think of this as the same factorisation. We note that a and au are associates, as are b and vb. So we think of two factorisations as the same if the factors in one are associates of factors in the other. For the same reason, we do not regard units as counting towards a factorisation, or we could multiply them *ad infinitum*.

This leads us to the appropriate definitions.

Definition Let R be an integral domain.

- An element $p \in R$ is **irreducible** if p is not zero or a unit, and if, whenever $p = ab$, either a or b is a unit (and the other is an associate of p).
- R is a **unique factorisation domain** or **UFD** if it holds that
  (a) every element other than zero and units can be factorised into irreducibles;
  (b) if $p_1 \cdots p_m = q_1 \cdots q_n$, where the p_i and q_j are irreducibles, then $m = n$, and (possibly after re-ordering) p_i and q_i are associates for $i = 1, \ldots, n$.

Briefly, condition (a) says that factorisations into irreducibles exist, while (b) says that they are 'unique up to order and associates'. (Note that any associate of an irreducible is irreducible.)

So the 'Fundamental Theorem of Arithmetic' says that $\mathbb{Z}$ is a UFD. (We will prove this later, in Section 2.13.) However, things here are a little different

from our first view of the FTA. Instead of factorising positive integers into positive primes, we factorise arbitrary integers into arbitrary (positive or negative) primes—remember that p and $-p$ are associates.

In a trivial way, a field is a UFD: it does not have any elements which are not zero or units!

One of the most substantial results about UFDs is the following:

Theorem 2.16 (Gauss' Lemma) *If R is a UFD, then $R[x]$ is a UFD.*

The proof of this result will be given in Chapter 7.

In particular, if F is a field, then $F[x]$ is a UFD. This will be proved in Section 2.13, since it uses techniques similar to those for $\mathbb{Z}$.

An important property of UFDs is that 'greatest common divisors exist'. In the case of the integers, we interpret the word 'greatest' in its usual sense for numbers. In general, this is not possible; the greatest common divisor is a common divisor which is divisible by every common divisor, in a sense which the next definition makes precise.

Definition Let R be a commutative ring.

- For $a, b \in R$, we say that a **divides** b (in symbols, $a \mid b$) if $b = ac$ for some $c \in R$.
- The element d is a **greatest common divisor** or **g.c.d.** of a and b if
  (a) d divides a and d divides b;
  (b) for any $e \in R$, if e divides a and e divides b then e divides d.

Thus, the greatest common divisor is not necessarily greatest in any absolute sense. In an arbitrary ring, two elements may have no greatest common divisor at all.

Theorem 2.17 *(a) In an integral domain, if a divides b and b divides a, then a and b are associates.*
(b) In an integral domain, if a and b have a greatest common divisor, then any two g.c.ds are associates.
(c) In a unique factorisation domain, every two elements have a greatest common divisor.

Proof (a) If $a = 0$ then $b = 0$ and there is nothing to prove. So suppose not. Let $b = ac$ and $a = bd$, then $a = acd$, so $a(1 - cd) = 0$. Since $a \neq 0$ and R is an integral domain, $cd = 1$; so c and d are units, and a and b are associates.

(b) If d_1 and d_2 are both g.c.ds of a and b, then (by part (b) of the definition) each divides the other; so they are associates.

(c) Assume that a and b are non-zero and not units. (Can you deal with the remaining cases?) Factorise a into irreducibles. Then, up to associates, every divisor of a is a product of some of the irreducibles in the factorisation of a. (For suppose that $a = xy$. Factorise x and y into irreducibles. Combining these gives a factorisation of a, which must be equal to the given one, up to order and associates.) So we find the g.c.d. of a and b by factorising both elements

into irreducibles, and taking all the irreducibles which (up to associates) occur in both factorisations. □

If this sounds somewhat complicated, it is just a generalisation of the argument which says, for example, that the g.c.d. of $2^4.3^2.5^2.7$ and $2^3.3^3.5.11$ is $2^3.3^2.5$.

This method finds the greatest common divisor, in principle. But it is not really an algorithm, since it depends on finding the factorisations of a and b, and we do not know how to do this in an arbitrary UFD (only that it can be done).

We conclude this section with an example of failure of the unique factorisation property.

Example Let $R = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$. Then R is a ring, with the usual definition of addition and multiplication of complex numbers. Moreover, R is an integral domain.

We first find the units of R. Let $a + b\sqrt{-5}$ be a unit; suppose that

$$(a + b\sqrt{-5})(x + y\sqrt{-5}) = 1.$$

Taking the square of the modulus of this equation (and using the fact that $|a + b\sqrt{-5}|^2 = a^2 + 5b^2$), we obtain

$$(a^2 + 5b^2)(x^2 + 5y^2) = 1.$$

Since a, b, x, y are integers, the only possibility is $b = y = 0$, $a^2 = x^2 = 1$, so that $a = \pm 1$. So the units are 1 and -1, and the associates of an element r are r and $-r$.

Consider the equation

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

We claim that all of the factors $2, 3, 1 + \sqrt{-5}, 1 - \sqrt{-5}$ are irreducible. Then certainly the factorisations are not the same up to order and associates!

To show that 2 is irreducible, suppose that

$$2 = (a + b\sqrt{-5})(x + y\sqrt{-5}).$$

Taking the norm squared as before, we obtain

$$4 = (a^2 + 5b^2)(x^2 + 5y^2).$$

As before, this implies that $b = y = 0$, so that $a = \pm 1$ or ± 2, and $x = \pm 2$ or ± 1. So one factor is a unit, and the other is an associate of 2. So 2 is irreducible. By a very similar argument, all the other factors are irreducible too.

So R is not a unique factorisation domain.

2.12 Principal ideal domains or PIDs. In the ring $\mathbb{Z}$, every ideal consists of all multiples of a fixed integer. This is a very important property, which we now study in general.

Definition Let $a_1, \ldots, a_n$ be elements of a ring R. The ideal **generated** by $a_1, \ldots, a_n$, denoted by $\langle a_1, \ldots, a_n \rangle$, is the smallest ideal containing these elements. ('Smallest' has the sense of inclusion; it is a subset of every ideal that contains $a_1, \ldots, a_n$.) Be aware that this is often written as $(a_1, \ldots, a_n)$. However, this risks confusion with the n-tuple $(a_1, \ldots, a_n)$. Angle brackets as used here are very common in mathematics to convey the idea of generation.

From the definition, it is not obvious that such an ideal exists. It can be shown that it does exist in any ring. But in a special case, it is easy to describe:

Proposition 2.18 *Let R be a commutative ring with identity, and let $a_1, \ldots, a_n$ be elements of R. Then*

$$\langle a_1, \ldots, a_n \rangle = \{x_1 a_1 + x_2 a_2 + \cdots + x_n a_n : x_1, \ldots, x_n \in R\}.$$

Proof Let $I = \{x_1 a_1 + \cdots + x_n a_n : x_1, \ldots, x_n \in R\}$, the set of all linear combinations of $a_1, \ldots, a_n$. We have to show that I is an ideal, that I contains $a_1, \ldots, a_n$, and that any ideal containing $a_1, \ldots, a_n$ necessarily contains all of I.

I is an ideal: (a) if $a, b \in I$, say $a = x_1 a_1 + \cdots + x_n a_n$ and $b = y_1 a_1 + \cdots + y_n a_n$, then

$$a - b = (x_1 - y_1)a_1 + \cdots + (x_n - y_n)a_n \in I.$$

(b) If $a = x_1 a_1 + \cdots + x_n a_n \in I$ and $r \in R$, then

$$ar = ra = (rx_1)a_1 + \cdots + (rx_n)a_n \in I.$$

So I passes the Ideal Test.

I contains $a_1, \ldots, a_n$: for $1 \leq i \leq n$, we have

$$a_i = 0a_1 + \cdots + 0a_{i-1} + 1a_i + 0a_{i+1} + \cdots + 0a_n \in I.$$

Any ideal containing $a_1, \ldots, a_n$ contains I: Let J be an ideal of R containing $a_1, \ldots, a_n$. For any $x_1, \ldots, x_n \in R$, we have $x_i a_i \in J$, and hence $x_1 a_1 + \cdots + x_n a_n \in J$ (using the fact that J is an ideal, and so is closed under addition and under multiplication by elements of R). So every element of I is in J, which means that $I \subseteq J$. □

An ideal of R is **principal** if it is generated by a single element. By the proposition, if R is a commutative ring with identity, then a principal ideal is of the form $\langle a \rangle = aR = \{ax : x \in R\}$. In other words, $\langle a \rangle$ consists of all elements divisible by a. In an integral domain, principal ideals have some further nice properties:

Proposition 2.19 *Let R be an integral domain.*

(a) For $a, b \in R$, if $\langle a \rangle = \langle b \rangle$, then a and b are associates.
(b) For $a, b \in R$, if $\langle a, b \rangle = \langle d \rangle$, then d is a greatest common divisor of a and b.

Proof (a) Suppose that $\langle a \rangle = \langle b \rangle$. Then $a \in \langle b \rangle = bR$, so b divides a, and similarly a divides b. By Theorem 2.17, a and b are associates.

(b) Suppose that $\langle a, b \rangle = \langle d \rangle$. Then $a, b \in \langle d \rangle$, so (as above) d divides both a and b. On the other hand, $d \in \langle a, b \rangle$, so $d = ax + by$ for some $x, y \in R$. Let e be any common divisor of a and b. Then $a = eu$ and $b = ev$ for some u, v. Then we have $d = ax + by = eux + evy = e(ux + vy)$; that is, e divides d. So d is a greatest common divisor. □

Definition A principal ideal domain or **PID** is an integral domain with the property that every ideal is principal.

Proposition 2.20 *(a) Let R be a PID. Then any two elements $a, b \in R$ have a greatest common divisor d, which can be written in the form $d = ax + by$ for some $x, y \in R$.*
(b) $\mathbb{Z}$ is a PID.

Proof (a) The ideal $\langle a, b \rangle$ is principal, and hence has the form $\langle d \rangle$ for some d. Now apply part (b) of the previous proposition.

(b) We found that every ideal of $\mathbb{Z}$ has the form $n\mathbb{Z}$, that is, $\langle n \rangle$ in the present notation. □

What has all this to do with factorisation? The following important result holds:

Theorem 2.21 *Every principal ideal domain is a unique factorisation domain.*

Proof Let R be a PID. We have to show two things: that elements of R can be factorised into irreducibles; and that the factorisation of an element is unique (up to order and associates). The first part is quite substantial; the proof is deferred until Chapter 7. I will show here that factorisations are unique. This depends on the following fact.

Proposition 2.22 *Let p be an irreducible element in a PID R. If p divides ab, then p divides a or p divides b.*

Proof Suppose that p divides ab but p does not divide a. Then the greatest common divisor of p and a is 1. (Remember that g.c.ds exist in a PID.) So there exist $x, y \in R$ such that $px + ay = 1$. Multiplying this equation by b, we obtain $pxb + aby = b$. Now p clearly divides pxb; and p divides aby (since p divides ab by assumption); so p divides b. The result is proved. □

It follows that if p is irreducible and p divides $a_1 \cdots a_n$, then p divides a_i for some i.

This property fails in the ring $R = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$ discussed in the preceding section. For 2 divides $6 = (1 + \sqrt{-5})(1 - \sqrt{-5})$, but 2 does not divide either factor.

Now we return to the proof that a PID is a UFD. Suppose that we have two factorisations of an element of the PID R, say

$$a = p_1 p_2 \cdots p_m = q_1 q_2 \cdots q_n,$$

where the p_i and q_j are irreducible. Now p_1 divides $q_1 \cdots q_n$, so by the remark following the proposition, p_1 divides q_j for some j. By re-ordering the product, we can assume that p_1 divides q_1. Since p_1 and q_1 are irreducible, they must be associates, say $p_1 = q_1 u$ for some unit u. Then we have

$$p_1 p_2 \cdots p_m = (q_1 u)(u^{-1} q_2) \cdots q_n.$$

By the Cancellation Law,

$$p_2 \cdots p_m = q_2' \cdots q_n,$$

where $q_2' = u^{-1} q_2$, an associate of q_2. Continuing in this manner we find that $m = n$ and that p_i and q_i are associates for all i (after suitable re-ordering), and we are done. □

We close this section with an example of a UFD which is not a PID. This is the ring $\mathbb{Z}[x]$ of polynomials over the integers. It is a UFD by Gauss' Lemma. The g.c.d. of the elements 2 and x is obviously 1; but there do not exist polynomials f and g such that $2f(x) + xg(x) = 1$, since the constant term of the left-hand side is even. Said otherwise, the ideal $\langle 2, x \rangle$ generated by 2 and x (which is the set of all polynomials whose constant term is even) cannot be generated by a single element.

2.13 Euclidean domains or EDs. We now look at an even more specialised class of rings (which, however, includes our prototype $\mathbb{Z}$ as well as polynomial rings over fields).

Definition Let R be a commutative ring with identity.

- A **Euclidean function** on R is a function d from the set of non-zero elements of R to the non-negative integers which satisfies
  (a) $d(ab) \geq d(a)$ for $a, b \neq 0$;
  (b) if $a, b \in R$ with $b \neq 0$, then there exist $q, r \in R$ with $a = bq + r$ and either $r = 0$ or $d(r) < d(b)$.
- R is a **Euclidean domain**, if there exists a Euclidean function on R.

Examples 1. $\mathbb{Z}$ is a ED. Take $d(a) = |a|$ for non-zero $a \in \mathbb{Z}$. If also $b \neq 0$, then clearly

$$d(ab) = |ab| = |a| \cdot |b| = d(a)d(b) \geq d(a),$$

since $d(b) \geq 1$. For (b), suppose that $b \neq 0$. If $b > 0$, divide a by b to obtain a quotient q and remainder r; then $a = bq + r$ with $0 \leq r < b$; that is, $r = 0$ or $d(r) < d(b) = b$. If $b < 0$, divide a by $-b$ instead.

2. For any field F, the polynomial ring $F[x]$ is a ED. In this case, we take the Euclidean function to be $d(f) = \deg(f)$, the degree of the polynomial f. (Recall that we did not define $\deg(f)$ if $f = 0$, but we do not need a value for $d(0)$ either.)

(a) If f, g are non-zero, then $d(fg) = d(f) + d(g) \geq d(f)$.

(b) Suppose that $g \neq 0$; we wish to find q, r with $f = gq + r$ and $r = 0$ or $\deg(r) < \deg(g)$. The proof is by induction on the degree of f. Let $m = \deg(f)$, $n = \deg(g)$. If $m < n$, we can take $q = 0$ and $r = f$. So suppose that $m \geq n$. Let

$$f(x) = a_m x^m + \text{ lower terms},$$
$$g(x) = b_n x^n + \text{ lower terms},$$

where a_m and b_n are non-zero. Put

$$f_1(x) = f(x) - (a_m b_n^{-1}) x^{m-n} g(x).$$

(This is defined since the coefficients form a field and $b_n \neq 0$.) The coefficient of x^m in f_1 is $a_m - (a_m b_n^{-1}) b_n = 0$, and clearly there are no terms of higher degree. So $\deg(f_1) < m = \deg(f)$. By the induction hypothesis, we have $f_1 = gq_1 + r_1$, where $r_1 = 0$ or $\deg(r_1) < \deg(g)$. Then

$$f = g(a_m b_n^{-1} x^{m-n} + q_1) + r_1;$$

so we can take $q = a_m b_n^{-1} x^{m-n} + q_1$, $r = r_1$.

The reason for the term 'Euclidean domain' is that this is the class of rings in which the Euclidean Algorithm for finding the greatest common divisor of two elements can be made to work. We met the Euclidean Algorithm for integers in Chapter 1. The general case is exactly the same. I will present it here in a way influenced by computer programming, as a recursive algorithm. But you do not need to know anything about computers in order to follow this. Just remember that an algorithm takes some data as input and produces some other data as output; we must specify what the algorithm is expected to do, and then we must prove that the algorithm really does what is claimed.

Euclidean Algorithm Let R be a Euclidean domain.

`Input:` Two elements $a, b \in R$.

`Output:` An element $c \in R$ which is a greatest common divisor of a and b. We write $c = \gcd(a, b)$ for this output.

`Operation:` If $b = 0$, then set $\gcd(a, b) = a$.

Otherwise, choose $q, r \in R$ such that $a = bq + r$ with either $r = 0$ or $d(r) < d(b)$; set $\gcd(a, b) = \gcd(b, r)$.

It is not clear that we have defined anything: why should $\gcd(b, r)$ be easier to calculate than $\gcd(a, b)$? Imagine that we are given two elements a and b, and are trying to find $\gcd(a, b)$. If $b = 0$, we obtain immediately the result a. Suppose not. Now observe that either $r = 0$, in which case we finish at the second step, or we have to calculate $\gcd(b, r)$ with $d(r) < d(b)$. During the calculation, the second alternative can only occur finitely often, since the value of d on the second argument of the function is strictly smaller at each instance, and a strictly decreasing sequence of non-negative integers cannot continue for ever. So, after a finite number of steps, the algorithm does terminate and produce a result.

Now, we have to show that it gives the correct result. The proof is by induction on $d(b)$ (taking the base case of the induction to be $b = 0$). In order to do this, we need to show two things:

(a) the greatest common divisor of a and 0 is a;
(b) if $a = bq + r$, then the greatest common divisor of a and b is equal to the greatest common divisor of b and r (up to associates).

The first fact should be clear if $a \neq 0$; you should think about it and convince yourself that it also holds if $a = 0$. (Use the definition of greatest common divisor, rather than any prejudices about greatest integers, etc.)

For the second point, observe that any divisor of a and b also divides $r = a - bq$, while any divisor of b and r also divides $a = bq + r$. So the set of all common divisors of a and b is the same as the set of common divisors of b and r, and the greatest common divisors must be associates, as required.

As we saw in case of the integers, the Euclidean Algorithm has another feature; it can be used to express the g.c.d. of a and b as a linear combination of these two elements. This is also true in general:

Enriched Euclidean Algorithm Let R be a Euclidean domain.

`Input:` Two elements $a, b \in R$.

`Output:` An element $c \in R$ which is a greatest common divisor of a and b, together with two elements x and y such that $c = ax + by$.

`Operation:` If $b = 0$, then we set $c = a$, $x = 1$, $y = 0$.

Otherwise, we write $a = bq + r$ with $r = 0$ or $d(r) < d(b)$ as usual, and apply the algorithm to b and r. Suppose that the output is c', x', and y'. Then we put $c = c'$, $x = y'$, and $y = x' - y'q$.

The proof that this algorithm terminates, and finds the g.c.d. correctly, is exactly as for the original version. We have to show that $c = xa + yb$. This is obvious in the first case of the algorithm. In the second case (arguing, as before, by induction), we may assume that $c' = bx' + ry'$. Then

$$c = c' = bx' + (a - bq)y' = ay' + b(x' - qy'),$$

as required.

Now we turn to some theoretical properties of Euclidean domains.

Proposition 2.23 *(a) A Euclidean domain is an integral domain.*

(b) If R is a Euclidean domain and a, b are non-zero elements with $a \mid b$ and $d(b) = d(a)$, then a and b are associates.

Proof (a) If a and b are non-zero, then $d(ab) \geq d(a)$, so $ab \neq 0$.

(b) By condition (b), $a = bq + r$, where $r = 0$ or $d(r) < d(a)$. Suppose that $r \neq 0$. Then a divides b, so a divides $r = a - bq$; by condition (a), $d(r) \geq d(a)$, a contradiction. So we must have $r = 0$, whence $a = bq$. So each of a and b divides the other, and these elements are associates. □

Now we come to the main result:

Theorem 2.24 *(a) A Euclidean domain is a principal ideal domain.*
(b) A Euclidean domain is a unique factorisation domain.

Proof (a) Let R be a Euclidean domain. Take any ideal $I \in R$: we have to show that I is a principal ideal. The argument is similar to our determination of the ideals in $\mathbb{Z}$. First, $0 \in I$; and, if I consists only of 0, then $I = \langle 0 \rangle$, and so I is principal. So we may suppose that I contains some non-zero elements. Choose an element $a \in I$ such that $d(a)$ is as small as possible. (This depends on the fact that the values of d are non-negative integers, so there is necessarily a smallest one.) We claim that $I = \langle a \rangle$. As usual, we have to show that any element of either set is contained in the other. First, take $x \in \langle a \rangle$; then x is of the form $x = ar$ for some $r \in R$, and so $x \in I$ (since $a \in I$ and I is an ideal). Conversely, take $x \in I$. By part (b) of the definition of a Euclidean function, we write $x = aq + r$, where $r = 0$ or $d(r) < d(a)$. Now $x \in I$ and $aq \in I$, so $r = x - aq \in I$; and, since a was chosen as an element of I with $d(a)$ as small as possible, it cannot happen that $d(r) < d(a)$, so we must have $r = 0$, and $x = aq \in \langle a \rangle$. Thus indeed $I = \langle a \rangle$.

(b) If we had proved Theorem 2.21, this would immediately follow from (a). Since we didn't do that, we have some work to do. We showed that factorisation is unique (up to order and associates) in any PID; so we only have to do the other part, to prove that any element (other than zero and units) has a factorisation in R. So take $a \in R$ with $a \neq 0$ and a not a unit. We show by induction on $d(a)$ that a has a factorisation. In other words, we assume that any element b with $d(b) < d(a)$ has a factorisation.

If a is irreducible, then we have a factorisation (with only one factor!), so suppose that $a = bc$, where neither b nor c is a unit. Now by condition (a), $d(b) \leq d(a)$ and $d(c) \leq d(a)$. If $d(b) < d(a)$ and $d(c) < d(a)$, then by the inductive hypothesis, both b and c have factorisations; combining these gives a factorisation of a. So we can suppose that $d(b) = d(a)$. But then a and b are associates, by part (b) of the Proposition: a contradiction. The proof is finished. □

Let us summarise our findings. We have three classes of integral domains:

- unique factorisation domains;
- principal ideal domains;
- Euclidean domains.

Theorems 2.21 and 2.24 show that

$$\text{ED} \Rightarrow \text{PID} \Rightarrow \text{UFD}.$$

We see the increasing strength of these conditions by looking at the facts about greatest common divisors:

- In a UFD, any two elements have a g.c.d.;
- In a PID, any two elements a, b have a g.c.d. d, and $d = xa + yb$ for some x, y;
- In a ED, any two elements a, b have a g.c.d. d, and $d = xa + yb$ for some x, y; moreover, d, x, y can be found by using the Euclidean Algorithm.

You might expect here an example of a PID which is not a ED. Such rings do exist; T. S. Motzkin showed that the ring

$$R = \{a + b\sqrt{-19} : \text{ either } a, b \in \mathbb{Z} \text{ or } a - \tfrac{1}{2}, b - \tfrac{1}{2} \in \mathbb{Z}\}$$

is an example. But the proof is more difficult. (You can read it in volume 55 of the *Bulletin of the American Mathematical Society*, starting on page 1142.)

Exercise 2.22 Show that, in a commutative ring with identity, no element can be both a zero-divisor and a unit.

Exercise 2.23 Write down the associate classes in the ring $\mathbb{Z}_{12}$.

Exercise 2.24 (∗) Find all positive integers m with the property that every unit a in $\mathbb{Z}_m$ satisfies $a^2 = 1$.

Exercise 2.25 Let R be an integral domain. Show that the units of $R[x]$ are precisely the constant polynomials which are units of R.

Exercise 2.26 (a) Let F be a field. Show that a matrix $A \in M_n(F)$ is a unit if and only if $\det(A) \neq 0$; and that a non-zero matrix A is a zero-divisor if and only if $\det(A) = 0$.

(∗) (b) More generally, let R be a commutative ring with identity. Prove that $A \in M_n(R)$ is a unit if and only if $\det(A)$ is a unit in R. [*Hint*: If $\det(A)$ is a unit, use the 'cofactor formula' to find an inverse of A. Conversely, if $AB = BA = I$, take determinants to show that $\det(A)$ is a unit.]

Exercise 2.27 Let x be an element in a ring with identity, and suppose that $x^n = 0$ for some positive integer n. Prove that $1+x$ is a unit. [*Hint*: $(1+x)(1-x+x^2-x^3+\cdots) = 1$.]

Exercise 2.28 (a) Find the greatest common divisor of the real polynomials $f(x) = x^2 + 3x + 2$ and $g(x) = x^5 + 2x^4 + 5x^3 + 6x + 2$.

(b) Give a simple description of the ideal $\langle f(x), g(x)\rangle$ of $\mathbb{R}[x]$.

Exercise 2.29 (a) Show that the ring $R = \{x + y\mathrm{i} : a, b \in \mathbb{Z}\}$ of Gaussian integers is a Euclidean domain, with Euclidean function $d(x + y\mathrm{i}) = x^2 + y^2$. [*Hint:* For (b), take

$a, b \in R$ with $b \neq 0$, and write $a/b = u + v\mathrm{i}$ in $\mathbb{C}$, where u and v are rational numbers. Then choose $m, n \in \mathbb{Z}$ such that

$$|(u + v\mathrm{i}) - (m + n\mathrm{i})| \leq 1/\sqrt{2},$$

by considering lattice points with integer coordinates in the complex plane.]

(∗) (b) Show that the ring of Eisenstein integers (Exercise 2.16(b)) is a Euclidean domain, by using the triangular lattice similarly.

Exercise 2.30 (a) Describe the units in the ring of Gaussian integers.

(∗) (b) Show that the irreducibles in the ring of Gaussian integers are of two types: primes $p \in \mathbb{Z}$ which cannot be written as the sum of two integer squares; and elements $x + y\mathrm{i}$, where $x^2 + y^2$ is a prime in $\mathbb{Z}$. (For example, 3 is a 'Gaussian prime'; 5 is not, since $5 = (2 + \mathrm{i})(2 - \mathrm{i})$, but the two factors are both Gaussian primes.)

Remark A theorem of Number Theory asserts that a prime p can be expressed as the sum of two squares if and only if $p = 2$ or $p \equiv_4 1$.

Fields

Recall that a field is a commutative ring with identity in which division is possible by non-zero elements. Rings are easy to build: we have seen polynomial rings, matrix rings, Boolean rings, cartesian products. Almost always, these rings turn out not to be fields. In fact, there are only two standard methods of constructing fields: applied to the integers, they produce the rationals, and the integers mod p.

2.14 Field of fractions. The first method involves going from a ring to its 'field of fractions', which generalises the construction of the rational numbers from the integers.

Let R be an integral domain. A field F is a **field of fractions** of R if

(a) R is a subring of F;
(b) Any element of F can be written in the form ab^{-1} for some $a, b \in R$ (where b^{-1} is calculated in F).

For example, the rational numbers $\mathbb{Q}$ form the field of fractions of the integers $\mathbb{Z}$. Any field is its own field of fractions.

Theorem 2.25 *Any integral domain has a field of fractions.*

Proof Let R be an integral domain. We let S be the set of all ordered pairs (a, b) for $a, b \in R$, $b \neq 0$. We intend that the ordered pair (a, b) will represent the element ab^{-1}. But, of course, $ab^{-1} = cd^{-1}$ if (and only if) $ad = bc$; so we want the ordered pairs (a, b) and (c, d) to represent the same element of F if this condition holds. Accordingly, we define an equivalence relation $\sim$ on S by the rule

$$(a, b) \sim (c, d) \text{ if and only if } ad = bc.$$

We prove that this really is an equivalence relation. First, $(a,b) \sim (a,b)$, since $ab = ba$ (R is commutative). Then, if $(a,b) \sim (c,d)$, then $ad = bc$, and so $cb = da$; this means $(c,d) \sim (a,b)$. Finally, suppose that $(a,b) \sim (c,d)$ and $(c,d) \sim (e,f)$. Then $ad = bc$ and $cf = de$. So

$$adf = bcf = bde,$$

and by the cancellation law (since $d \neq 0$) we deduce that $af = be$; so $(a,b) \sim (e,f)$. So $\sim$ really is an equivalence relation.

Now we let $[a,b]$ denote the equivalence class of the ordered pair (a,b): $[a,b] = \{(c,d) : ad = bc\}$. Let F be the set of equivalence classes. We define addition and multiplication on F by the following rules:

$$[a,b] + [c,d] = [ad + bc, bd],$$
$$[a,b] \cdot [c,d] = [ac, bd].$$

(To see where these definitions come from, work out what you would expect $(ab^{-1}) + (cd^{-1})$ and $(ab^{-1})(cd^{-1})$ to be, by the usual rules for fractions.)

We have to check that these operations are well defined, and that they do indeed make F a field. All of this is straightforward checking. Finally, the map that takes a to the equivalence class $[a,1]$ is a one-to-one homomorphism from R to F, so we can regard R as a subring of F. Moreover, the inverse of the element $[b,1]$ is $[1,b]$ if $b \neq 0$; and $[a,b] = [a,1][b,1]^{-1}$. So, if we identify R with its image in F under this embedding, we see that F is indeed a field of fractions of R. □

In fact, the field of fractions is unique (up to isomorphism). This is another way of saying that the only possible way to construct a field of fractions is the way we actually did it.

2.15 Maximal ideals and fields. The second method of constructing fields generalises the passage from the integers to the integers modulo a prime: the field is constructed as a factor ring. To study this, first we need a different test for when a ring is a field.

Proposition 2.26 *Let R be a commutative ring with identity. Then R is a field if and only if the only ideals in R are $\{0\}$ and R itself.*

Proof For the forward implication, suppose that R is a field. Take any ideal I of R. Suppose that $I \neq \{0\}$; we have to show that $I = R$, that is, that every element of R is in I. Certainly, some non-zero element is in I, say $a \in I$. Now, for any $x \in R$, we have $x = (xa^{-1})a \in I$. So $I = R$.

For the converse, let R be a commutative ring with identity whose only ideals are $\{0\}$ and R. We have to show that all its non-zero elements have inverses. So take $a \in R$ with $a \neq 0$. Let I be the ideal $(a) = aR$. Then $I \neq \{0\}$, since $a \in I$; so $I = R$. Thus, $1 \in I = aR$, so there exists $b \in R$ with $ab = ba = 1$, as required. □

We say that the ideal I of the ring R is a **maximal ideal** of R if $I \neq R$ but there is no ideal J properly between I and R; that is, if J is an ideal with $I \subseteq J \subseteq R$, then $J = I$ or $J = R$.

Theorem 2.27 *Let R be a commutative ring with identity, and I an ideal of R. Then R/I is a field if and only if I is a maximal ideal of R.*

Proof This follows immediately from the proposition and the correspondence between ideals of R/I and ideals of R containing I given by the Second Isomorphism Theorem (Theorem 2.11). □

How do we recognise maximal ideals?

Proposition 2.28 *Let R be a principal ideal domain, and take $a \in R$ with $a \neq 0$. Then $\langle a \rangle$ is a maximal ideal of R if and only if a is irreducible.*

Proof In an integral domain, we have $\langle a \rangle \subseteq \langle b \rangle$ if and only if b divides a, and $\langle a \rangle = \langle b \rangle$ if and only if a and b are associates. In particular, $\langle a \rangle = R$ if and only if a is a unit (an associate of 1: note that $\langle 1 \rangle = R$). Hence $\langle a \rangle$ is maximal if and only if every element b which divides a is either an associate of a or a unit; but this is exactly the condition that a is irreducible. Finally, if R is a principal ideal domain, then there are no other ideals to spoil the maximality of (a). □

Example $R = \mathbb{Z}$. We see that $\mathbb{Z}_n$ is a field if and only if n is prime. (In fact we knew this already. For we showed that m is a unit in $\mathbb{Z}_n$ if and only if m and n are coprime; and this holds for all non-zero residues mod n if and only if n is prime.)

We will apply this result to polynomial rings in the next section.

2.16 Field extensions, finite fields. The standard procedure for constructing the complex numbers from the real numbers is to 'adjoin' a square root of -1; that is, an element i satisfying $\mathrm{i}^2 + 1 = 0$. We will now describe this procedure, 'adjoining the root of a polynomial', in more detail.

Theorem 2.29 *Let F be a field, and f a polynomial which is irreducible in $F[x]$. Then there is a field K containing F and an element α satisfying $f(\alpha) = 0$.*

Proof The construction is simple. We set $K = F[x]/\langle f \rangle$. This is a field by the results of the last section: $F[x]$ is a principal ideal domain and we are given that f is irreducible, so $\langle f \rangle$ is a maximal ideal in $F[x]$, and $F[x]/\langle f \rangle$ is a field.

We have to show that

(a) K contains (a field isomorphic to) F;
(b) K contains a root α of f.

(a) Set $I = \langle f \rangle$. For $a \in F$, let $\overline{a}$ denote the coset $I + a$, and let $\overline{F}$ be the set of all such cosets. We show that the map $a \mapsto \overline{a}$ is an isomorphism from F to $\overline{F}$. It is one-to-one, since if $\overline{a} = \overline{b}$ then $b - a \in \langle f \rangle$, so $b - a = 0$ (as any non-zero

element of $\langle f\rangle$ has degree at least as great as that of f). The homomorphism property is clear.

(b) We take α to be the coset $I+x$. Let

$$f(x) = a_n x^n + \cdots + a_1 x + a_0.$$

Then

$$\begin{aligned} f(\alpha) &= \overline{a_n}(I+x)^n + \cdots + \overline{a_1}(I+x) + \overline{a_0}(I+1) \\ &= (I + a_n x^n) + \cdots + (I + a_1 x) + (I + a_0) \\ &= I + (a_n x^n + \cdots + a_1 x + a_0) \\ &= I + f(x) \\ &= I \end{aligned}$$

and we are done, since I is the zero element of $F[x]/I$. □

Our proof shows that a field with the required properties exists. However, it is defined as a factor ring, which is not the most convenient form for calculation. As always, calculation in a factor ring is very much easier if we can make a good choice of coset representatives!

Let f be an irreducible polynomial of degree $n > 0$ over the field F. We claim that every coset of the ideal $\langle f\rangle$ in $F[x]$ has a unique representative r satisfying $r = 0$ or $\deg(r) < n$. Such an r exists because of the Euclidean property of $F[x]$. (If g is any polynomial, and $g = fq + r$, then g and r differ by a multiple of f, and so $\langle f\rangle + g = \langle f\rangle + r$.) If r_1 and r_2 are two representatives of the same coset with $r_i = 0$ or $\deg(r_i) < n$ for $i = 1, 2$, then $\langle f\rangle + r_1 = \langle f\rangle + r_2$, so $r_1 - r_2 \in \langle f\rangle$; this means that f divides $r_1 - r_2$. But since $\deg(f) = n$, this implies that $r_1 - r_2 = 0$, so $r_1 = r_2$.

Moreover, a simple argument (similar to the one in the above proof) shows that the coset $\langle f\rangle + r$ is equal to $r(\alpha)$, where α is the coset $\langle f\rangle + x$.

This means that

> Every element of $K = F[x]/\langle f\rangle$ can be uniquely expressed in the form
> $$c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_{n-1}\alpha^{n-1},$$
> where $c_0, c_1, \ldots, c_{n-1} \in F$.
> The addition and multiplication in K are given by the usual arithmetic rules, with the added condition that $f(\alpha) = 0$.

The construction of $\mathbb{C} = \mathbb{R}[x]/\langle x^2 + 1\rangle$ is a familiar example: every complex number is uniquely expressible as $c_0 + c_1\mathrm{i}$, where $\mathrm{i}^2 = -1$.

For another example, let us construct a finite field of order 4.

We start with the field $F = \mathbb{Z}_2$, with elements 0 and 1. Consider the polynomial $x^2 + x + 1 \in F[x]$. This polynomial is irreducible, since the only possible factorisation would be into two linear factors, which would imply that the polynomial has a root in F; but $0^2 + 0 + 1 \neq 0$ and $1^2 + 1 + 1 \neq 0$. Let

$K = F[x]/\langle x^2+x+1\rangle$, and let α be the coset $\langle x^2+x+1\rangle+x$, so that $\alpha^2+\alpha+1 = 0$. The field K has four elements: $K = \{0, 1, \alpha, \alpha+1\}$. (This is our canonical representation above.) Letting $\beta = \alpha+1 = \alpha^2$ (noting that $x = -x$ in the field K), we obtain the following tables:

$+$	0	1	α	β
0	0	1	α	β
1	1	0	β	α
α	α	β	0	1
β	β	α	1	0

$\cdot$	0	1	α	β
0	0	0	0	0
1	0	1	α	β
α	0	α	β	1
β	0	β	1	α

If p is a prime, and n a positive integer, then a finite field of order p^n can be constructed in the same way if an irreducible polynomial of degree n over $\mathbb{Z}_p$ can be found. Galois showed that this is always possible:

Theorem 2.30 *For any prime number p and any positive integer n, there is an irreducible polynomial of degree n over $\mathbb{Z}_p$, and hence a finite field of order p^n.*

Exercise 2.31 Show that the polynomial $x^2 + 1$ is irreducible over $\mathbb{Z}_3$, and hence construct a field of order 9.

Exercise 2.32 Show that the polynomials x^3+x+1 and x^3+x^2+1 are both irreducible over $\mathbb{Z}_2$. Are the corresponding fields of order 8 isomorphic?

Exercise 2.33 Show that, if F is a field with q elements and f an irreducible polynomial of degree n over F, then the field $K = F[x]/\langle f\rangle$ has q^n elements.

Exercise 2.34 Prove that any two fields of fractions F_1 and F_2 of an integral domain R are isomorphic, where the isomorphism $\theta : F_1 \to F_2$ can be chosen so that its restriction to the subring R of F_1 is the identity map.

Exercise 2.35 A subset X of a ring R is called **multiplicatively closed** if $a, b \in X$ implies $ab \in X$.

(a) Prove that R is an integral domain if and only if the set of non-zero elements of R is multiplicatively closed.

*(b) Let R be a commutative ring with identity, and let X be a multiplicatively closed subset of R containing 1 but not 0. Define an equivalence relation $\sim$ on $R \times X$ by the rule that $(a, b) \sim (c, d)$ if and only if $ad = bc$. Define operations of addition and multiplication on the set F of equivalence classes of $\sim$ as in Section 2.14. Prove that F is a ring containing R, in which every element of X has an inverse, and every element of F can be written as ab^{-1}, where $a \in R$ and $b \in X$.

Appendix: Miscellany

We end with some miscellaneous topics.

2.17 Cage on zero. The American composer John Cage wrote the following. What is he talking about? (Think about this before reading the following

discussion.)

> Curiously enough, the twelve-tone system has no zero in it. Given a series: 3, 5, 2, 7, 10, 8, 11, 9, 1, 6, 4, 12 and the plan of obtaining its inversion by numbers which when added to the corresponding ones of the original series will give 12, one obtains 9, 7, 10, 5, 2, 4, 1, 3, 11, 6, 8 and 12. For in this system 12 plus 12 equals 12. There is not enough of zero in it.
>
> John Cage (1968).

I contend that Cage is confusing two different zeros, the zero element of the real numbers and the zero element of the integers mod 12.

Real numbers Cage was very much attracted to the Zen concept of emptiness. One of his most famous compositions, entitled $4'33''$, involves a pianist sitting at the keyboard of a piano for 4 minutes and 33 seconds without striking a note; the audience notices the background noise (since no emptiness is truly empty). The real numbers represent sound intensity, so zero is the absence of sound.

Integers mod 12 Musical notation is based on the fact that notes an octave apart (that is, when the frequency of one is double that of the other) have a very similar subjective effect in melodic terms. So we regard such notes as 'equivalent'. More generally, two notes are equivalent if they are a whole number of octaves apart.

In Western music, only a discrete set of notes is used. The octave is divided into 12 intervals called **semitones**. Thus, the semitones appear (on a keyboard, say), stretching to infinity in both directions like the integers. As above, two semitones are equivalent if they differ by a whole number of octaves; that is, if (as integers) they are congruent mod 12. So the musical scale, for thematic purposes, has the structure of the integers mod 12. Various musical operations fit into this framework. For example, transposition just involves adding a fixed constant to each note. Inversion involves replacing each equivalence class by its negative. (This is what Cage describes.)

Two kinds of zeros The equivalence classes referred to are the congruence classes mod 12, that is, the cosets of $12\mathbb{Z}$ in $\mathbb{Z}$. We can make any choice of coset representatives we like. Mathematicians usually use $0, 1, 2, \ldots, 11$. Musicians use 12 instead of 0 as the representative of the class $12\mathbb{Z}$, so that their semitones are labelled $1, 2, 3, \ldots, 12$.

Now Cage's arithmetic checks, since $-3 = 9$, $-5 = 7$, and so on, in $\mathbb{Z}_{12}$ (the integers mod 12). The mathematician says $-0 = 0$, the musician $-12 = 12$; it is exactly the same, just involving a different choice of coset representative.

So, contrary to what Cage says, there is a zero in the twelve-tone scale (but musicians call it 12); and it has nothing to do with the real number zero, the zero of intensity or absence of sound when the pianist is not striking the keys.

Footnote The title of Cage's piece mentioned above itself blurs the categories between different kinds of numbers. Four minutes and thirty-three seconds make 273 seconds; and -273 is the temperature of absolute zero in the Celsius scale. If he had used the Fahrenheit scale, Cage would presumably have titled his piece $7'39''$; but this duration may have taxed the patience of the audience too far!

2.18 Solution to Exercise 2.10. Exercise 2.10 asks whether it is possible to have a ring whose elements are the integers, with (a) the same addition as $\mathbb{Z}$ but different multiplication, or (b) the same multiplication but different addition.

Part (a) is easy if you remember the definition of a zero ring from Section 2.2: if we are given the operation of addition satisfying axioms (A0)–(A4), and we define multiplication by $ab = 0$ for all a, b, we obtain a ring. (For a more challenging question, try to describe all the possible definitions of multiplication which would give a ring.)

Part (b) is much harder. If you tried this question, you probably attempted to write down an explicit rule for addition which would make R into a ring. I do not know how to do that. Instead, I will give here a solution which is non-constructive, and is an illustration of the concept of factorisation, which we discussed in Sections 2.10–2.13.

Let R be a ring in which the set of elements is $\mathbb{Z}$ and the multiplication is the same as that in $\mathbb{Z}$. We start by making a list of properties of R. Any property which is defined purely in terms of multiplication, which holds in $\mathbb{Z}$, will hold in R. Thus, we have the following:

(a) R is commutative.
(b) R has an identity element 1.
(c) R has no divisors of zero. (Thus, R is an integral domain.)
(d) R has just two units, 1 and z, where $z^2 = 1$. (In fact, in $\mathbb{Z}$, we have $z = -1$, but -1 depends on the addition, so we cannot assert that $z = -1$ here. Confusingly, z is the integer whose name is -1, but we do not know that it is the additive inverse of 1.)
(e) R has infinitely many irreducibles (by the theorem of Euclid).
(f) R is a unique factorisation domain.

In fact, these properties determine the multiplication in R completely. For, if they hold, then any non-zero element can be uniquely written as $up_{k_1}^{a_1} \cdots p_{k_r}^{a_r}$, where $u = 1$ or z, $p_1, p_2, \ldots$ are the irreducibles (one from each associate class), and $a_1, \ldots, a_r$ are positive integers. Now the rule for multiplying these elements is clear.

This means that, if we can find a ring S different from $\mathbb{Z}$ having properties (a)–(f), then S will have the same multiplication as $\mathbb{Z}$, and so $R = S$ is a solution to the problem.

The simplest example of such a ring is the polynomial ring $F[x]$, where $F = \mathbb{Z}_3$ is the field of integers mod 3. This is a UFD (since it is a Euclidean domain); its units are the two non-zero constants; and Euclid's proof holds virtually unchanged to show that there are infinitely many irreducibles. (If there were

only finitely many irreducibles, say $f_1, \ldots, f_r$, then the polynomial $f_1 \cdots f_r + 1$ would not be irreducible, but would not be divisible by any irreducible, a contradiction.)

For example, we might let the irreducible polynomials x, $x+1$, $x-1$, x^2+1, ... correspond to the prime numbers 2, 3, 5, 7, ... Then, using $\oplus$ for the new addition, we have $1 \oplus 1 = -1$, $3 \oplus 5 = -2$, $4 \oplus 1 = 7$, $7 \oplus 1 = 15$, and so on.

2.19 Ideals in matrix rings. A commutative ring R with identity, whose only ideals are the trivial ones (namely, $\{0\}$ and R), is necessarily a field (see Proposition 2.24). This is false if we do not assume commutativity. The ring $M_n(F)$ of all $n \times n$ matrices over the field F has only the trivial ideals, as we shall see; but it is not a division ring for $n > 1$, since there are non-zero singular matrices.

Theorem 2.31 *Let R be a commutative ring with identity, and n a positive integer.*

(a) If S is an ideal of R, then $M_n(S)$ is an ideal of $M_n(R)$.
(b) Every ideal of $M_n(R)$ is of this form.

Proof (a) If S is an ideal of R, then it is a ring, and so $M_n(S)$ is a ring, so (by definition) a subring of $M_n(R)$. Now take $A = (a_{ij}) \in M_n(S)$, and $X = (x_{ij}) \in M_n(R)$. The (i,j) entry of AX is

$$\sum_{k=1}^{n} a_{ik} x_{kj}.$$

Now $a_{ik}x_{kj} \in S$, since $a_{ik} \in S$ and S is an ideal. Summing over k then gives an element of S. So $AX \in M_n(S)$. Similarly $XA \in M_n(S)$. So $M_n(S)$ is an ideal of $M_n(R)$.

(b) Suppose that T is an ideal of $M_n(R)$. Let S be the set of elements of R which occur as entries in matrices in T. We show that S is an ideal of R and that $T = M_n(S)$.

Let E_{ij} denote the matrix with 1 in row i and column j, and 0 in all other positions. Also, let S' be the set of all elements $x \in R$ such that $xE_{11} \in T$. (Here xE_{11} is the matrix with x in the top left-hand corner and all other entries zero.)

Step 1 $S' = S$.

For clearly $S' \subseteq S$. Let $x \in S$; then there is a matrix $A = (a_{ij}) \in T$ such that $a_{pq} = x$. Now it is easily checked that

$$E_{1p} A E_{q1} = a_{pq} E_{11} = xE_{11}.$$

Since T is an ideal, $xE_{11} \in T$, so $x \in S'$.

Step 2 S is an ideal of R.

Take $x, y \in S$ and $r \in R$. Then $xE_{11}, yE_{11} \in T$. Then, for any $r \in R$, we have

$$\begin{aligned}(x+y)E_{11} &= xE_{11} + yE_{11} \in T,\\ (rx)E_{11} &= (rE_{11})(xE_{11}) \in T,\\ (xr)E_{11} &= (xE_{11})(rE_{11}) \in T,\end{aligned}$$

since T is an ideal of $M_n(R)$. So $x+y, rx, xr \in S$, and S is an ideal of R.

Step 3 $T = M_n(S)$.

By definition, $T \subseteq M_n(S)$. Suppose that $A = (a_{ij}) \in M_n(S)$. Then

$$A = \sum_{i,j=1}^{n} E_{i1}(a_{ij}E_{11})E_{1j} \in T,$$

since $a_{ij}E_{11} \in T$ by Step 1 and T is an ideal. □

Thus, for example, the ring of 2×2 matrices over the ring of integers mod 4 has just three ideals:

- the zero ideal;
- the ideal consisting of matrices with every entry 0 or 2;
- the whole ring.

A ring R is defined to be **simple** if the only ideals in R are $\{0\}$ and R. Thus, any field (or, indeed, any division ring) is simple. From Theorem 2.31 we immediately conclude:

Corollary 2.32 *Let F be a field, and n a positive integer. Then $M_n(F)$ is a simple ring.*

Exercise 2.36 Recall the definition of the **direct product** $R \times S$ of two rings R and S: the elements of $R \times S$ are all ordered pairs (r, s), where $r \in R$ and $s \in S$; and the operations are componentwise, that is,

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2), \qquad (r_1, s_1) \cdot (r_2, s_2) = (r_1 r_2, s_1 s_2).$$

Let $R' = \{(r, 0) : r \in R\}$ and $S' = \{(0, s) : s \in S\}$. Prove that R' and S' are ideals of $R \times S$ isomorphic to R and S respectively.

Now suppose that T is a ring which contains ideals R and S having the property that every element of T can be written uniquely in the form $r + s$, where $r \in R$ and $s \in S$. Prove the following assertions:

(a) $R + S = T$ and $R \cap S = \{0\}$.
(b) If $r \in R$ and $s \in S$, then r and s commute (that is, $rs = sr$).
(c) The map θ from $R \times S$ to T given by $(r, s)\theta = r + s$ is an isomorphism.
(d) T is isomorphic to $R \times S$.

Exercise 2.37 Let R be a ring, and a an element of R. Remember that $\langle a\rangle$ means the **ideal generated by** a, which by definition is the smallest ideal of R containing a.

(a) Prove that $\langle a\rangle$ is the set of all elements of R of the form

$$na + sa + at + \sum_{i=1}^{m} s_i a t_i,$$

where m and n are integers, $m \geq 0$, and s, t, s_i, t_i are elements of R (and na has its usual meaning).

(b) Suppose that R has an identity. Show that the terms na, sa, and at can be dropped from the expression above.

(c) The element a is said to be **central** if it commutes with every element of R. Show that, if R has an identity and a is central, then

$$\langle a\rangle = aR = \{ar : r \in R\}.$$

(d) Give a description of $\langle a\rangle$ in the case where a is central but R does not necessarily have an identity.

Exercise 2.38 Let F be a field and n a positive integer. Let R be the ring $M_n(F)$ of $n \times n$ matrices over F. Let $a = E_{11}$ be the matrix with entry 1 in the first row and column, and all other entries zero. By Theorem 2.31, we know that $\langle a\rangle = R$. So, by part (b) of the preceding exercise, every element of R can be written in the form $\sum_{i=1}^{m} s_i a t_i$, for some elements $s_i, t_i \in R$. Show that there are elements of R which cannot be expressed as the sum of fewer than n terms of the form $s_i a t_i$, for $s_i, t_i \in R$.

Exercise 2.39 An element e of a ring is said to be an **idempotent** if $e^2 = e$.

(a) Let e be an idempotent in a ring with identity. Show that $1 - e$ is also an idempotent.

(b) Let R and S be rings with identity. Show that the elements $(1, 0)$ and $(0, 1)$ are central idempotents of the direct product $R \times S$, whose sum is the identity of $R \times S$.

(c) Conversely, suppose that T is a ring with identity and e is a central idempotent of T with $e \neq 0, 1$. Prove that $T \cong R \times S$, where $R = eT$ and $S = (1 - e)T$.

Exercise 2.40 An element r of a ring R is said to be **nilpotent** if $r^n = 0$ for some positive integer n.

(a) Prove that a non-zero nilpotent element is a zero divisor. Is the converse true?

(b) Prove that, in a commutative ring, the set of nilpotent elements is an ideal.

(c) Let n be a positive integer. Find all nilpotent elements of the ring $\mathbb{Z}_n$ of integers mod n. (Since, by the previous part they form an ideal, they must consist of all multiples of n^* for some integer n^* dividing n. Your job is to calculate n^* in terms of n.)

Exercise 2.41 Let R be a ring with identity. Prove that, if the element $r \in R$ is nilpotent, then $1 + r$ is a unit.

Exercise 2.42 Prove that, in a matrix ring $M_n(R)$, any strictly upper triangular matrix is nilpotent.