Rings

2.1 Introduction

A ring can be thought of as a generalisation of the integers, $\mathbb{Z}$. We can add and multiply elements of a ring, and we are interested in such questions as factorisation into primes, construction of "modular arithmetic", and so on.

2.1.1 Definition of a ring

Our first class of structures are *rings*. A ring has two operations: the first is called *addition* and is denoted by $+$ (with infix notation); the second is called *multiplication*, and is usually denoted by juxtaposition (but sometimes by $\cdot$ with infix notation).

In order to be a ring, the structure must satisfy certain rules called *axioms*. We group these into three classes. The name of the ring is R.

We define a *ring* to be a set R with two binary operations satisfying the following axioms:

Axioms for addition:

(A0) (*Closure law*) For any $a, b \in R$, we have $a+b \in R$.

(A1) (*Associative law*) For any $a, b, c \in R$, we have $(a+b)+c = a+(b+c)$.

(A2) (*Identity law*) There is an element $0 \in R$ with the property that $a+0 = 0+a = a$ for all $a \in R$. (The element 0 is called the *zero element* of R.)

(A3) (*Inverse law*) For any element $a \in R$, there is an element $b \in R$ satisfying $a+b = b+a = 0$. (We denote this element b by $-a$, and call it the *additive inverse* or *negative* of a.)

(A4) (*Commutative law*) For any $a, b \in R$, we have $a+b = b+a$.

Axioms for multiplication:

(M0) (*Closure law*) For any $a, b \in R$, we have $ab \in R$.

(M1) (*Associative law*) For any $a, b, c \in R$, we have $(ab)c = a(bc)$.

Mixed axiom:

(D) (*Distributive laws*) For any $a, b, c \in R$, we have $(a+b)c = ac+bc$ and $c(a+b) = ca+cb$.

Remarks 1. The closure laws (A0) and (M0) are not strictly necessary. If $+$ is a binary operation, then it is a function from $R \times R$ to R, and so certainly $a+b$ is an element of R for all $a, b \in R$. We keep these laws in our list as a reminder.

2. The zero element 0 defined by (A2) and the negative $-a$ defined by (A3) are not claimed to be unique by the axioms. We will see later on that there is only one zero element in a ring, and that each element has only one negative.

Axioms (M0) and (M1) parallel (A0) and (A1). Notice that we do not require multiplicative analogues of the other additive axioms. But there will obviously be some rings in which they hold. We state them here for reference.

Further multiplicative properties

(M2) (*Identity law*) There is an element $1 \in R$ such that $a1 = 1a = a$ for all $a \in R$. (The element 1 is called the *identity element* of R.)

(M3) (*Inverse law*) For any $a \in R$, if $a \neq 0$, then there exists an element $b \in R$ such that $ab = ba = 1$. (We denote this element b by a^{-1}, and call it the *multiplicative inverse* of a.)

(M4) (*Commutative law*) For all $a, b \in R$, we have $ab = ba$.

A ring which satisfies (M2) is called a *ring with identity*; a ring which satisfies (M2) and (M3) is called a *division ring*; and a ring which satisfies (M4) is called a *commutative ring*. (Note that the term "commutative ring" refers to the fact that the multiplication is commutative; the addition in a ring is always commutative!) A ring which satisfies all three further properties (that is, a commutative division ring) is called a *field*.

2.1.2 Examples of rings

1. The integers

The most important example of a ring is the set $\mathbb{Z}$ of integers, with the usual addition and multiplication. The various properties should be familiar to you; we will simply accept that they hold. $\mathbb{Z}$ is a commutative ring with identity. It is not a division ring because there is no *integer* b satisfying $2b = 1$. This ring will be our prototype for several things in the course.

Note that the set $\mathbb{N}$ of natural numbers, or non-negative integers, is not a ring, since it fails the inverse law for addition. (There is no non-negative integer b such that $2 + b = 0$.)

2. Other number systems

Several other familiar number systems, namely the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$, are fields. Again, these properties are assumed to be familiar to you.

3. The quaternions

There do exist division rings in which the multiplication is not commutative, that is, which are not fields, but they are not so easy to find. The simplest example is the ring of *quaternions*, discovered by Hamilton in 1843.

On 16 October 1843 (a Monday) Hamilton was walking in along the Royal Canal with his wife to preside at a Council meeting of the Royal Irish Academy. Although his wife talked to him now and again Hamilton hardly heard, for the discovery of the quaternions, the first noncommutative [ring] to be studied, was taking shape in his mind. He could not resist the impulse to carve the formulae for the quaternions in the stone of Broome Bridge (or Brougham Bridge as he called it) as he and his wife passed it.

Instead of adding just one element i to the real numbers, Hamilton added three. That is, a *quaternion* is an object of the form $a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$, where

$$\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1, \quad \mathrm{ij} = -\mathrm{ji} = \mathrm{k}, \quad \mathrm{jk} = -\mathrm{kj} = \mathrm{i}, \quad \mathrm{ki} = -\mathrm{ik} = \mathrm{j}.$$

It can be shown that all the axioms (A0)–(A4), (M0)–(M3) and (D) are satisfied.

For example, if a, b, c, d are not all zero, then we have

$$(a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}) \left(\frac{a - b\mathrm{i} - c\mathrm{j} - d\mathrm{k}}{a^2 + b^2 + c^2 + d^2} \right) = 1.$$

The ring of quaternions is denoted by $\mathbb{H}$, to commemorate Hamilton.

4. Matrix rings

We briefly defined addition and multiplication for matrices in the last chapter. The formulae for addition and multiplication of $n \times n$ matrices, namely

$$(A + B)_{ij} = A_{ij} + B_{ij}, \qquad (AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj},$$

just depend on the fact that we can add and multiply the entries. In principle these can be extended to any system in which addition and multiplication are possible. However, there is a problem with multiplication, because of the $\sum_{k=1}^{n}$, which tells us to add up n terms. In general we can only add two things at a time, since addition is a binary operation, so we have to make the convention that, for example, $a + b + c$ means $(a + b) + c$, $a + b + c + d$ means $(a + b + c) + d$, and so on. We will return to this point in the next subsection.

Now we have the following result:

Proposition 2.1 *Let R be a ring. Then the set $M_n(R)$ of $n \times n$ matrices over R, with addition and multiplication defined in the usual way, is a ring. If R has an identity, then $M_n(R)$ has an identity; but it is not in general a commutative ring or a division ring.*

We will look at the proof later, once we have considered addition of n terms.

5. Polynomial rings

In much the same way, the usual rules for addition of polynomials,

$$\left(\sum a_i x^i\right) + \left(\sum b_i x^i\right) = \sum (a_i + b_i) x^i, \qquad \left(\sum a_i x^i\right) \left(\sum b_i x^i\right) = \sum d_i x^i,$$

where

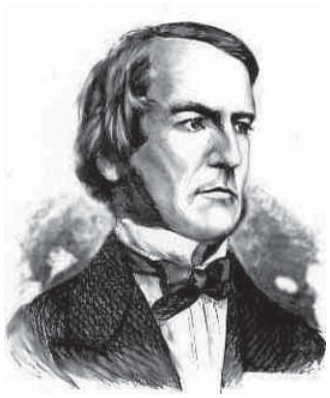
$$d_i = \sum_{k=0}^{i} a_k b_{i-k},$$

can be extended to polynomials with coefficients in any algebraic structure in which addition and multiplication are defined. As for matrices, we have to be able to add an arbitrary number of terms to make sense of the definition of multiplication. We have the result:

Proposition 2.2 *Let R be a ring, then the set $R[x]$ of polynomials over R, with addition and multiplication defined in the usual way, is a ring. If R is commutative, then so is $R[x]$; if R has an identity, then so does $R[x]$; but it is not a division ring.*

Again we defer looking at the proof.

6. Rings of sets



The idea of forming a ring from operations on sets is due to George Boole, who published in 1854 *An investigation into the Laws of Thought, on Which are founded the Mathematical Theories of Logic and Probabilities*. Boole approached logic in a new way reducing it to algebra, in much the same way as Descartes had reduced geometry to algebra.

The familiar set operations of union and intersection satisfy some but not all of the ring axioms. They are both commutative and associative, and satisfy the distributive laws both ways round; but they do not satisfy the identity and inverse laws for addition.

Boole's algebra of sets works as follows. Let $\mathscr{P}(A)$, the *power set* of A, be the set of all subsets of the set A. Now we define addition and multiplication on $\mathscr{P}(A)$ to be the operations of symmetric difference and intersection respectively:

$$x+y = x \triangle y, \qquad xy = x \cap y.$$

Proposition 2.3 *The set $\mathscr{P}(A)$, with the above operations, is a ring; it is commutative, has an identity element, but is not a field if $|A| > 1$. It satisfies the further conditions $x+x=0$ and $xx=x$ for all x.*

We won't give a complete proof, but note that the empty set is the zero element (since $x \triangle \emptyset = x$ for any set x), while the additive inverse $-x$ of x is equal to x itself (since $x \triangle x = \emptyset$ for any x). Check the other axioms for yourself with Venn diagrams.

A ring satisfying the further condition that $xx = x$ for all x is called a *Boolean ring*.

7. Zero rings

Suppose that we have any set R with a binary operation $+$ satisfying the additive axioms (A0)–(A4). (We will see later in the course that such a structure is called an *abelian group*.) Then we can make R into a ring by defining $xy = 0$ for all $x, y \in R$. This is not a very exciting rule for multiplication, but it is easy to check that all remaining axioms are satisfied.

A ring in which all products are zero is called a *zero ring*. It is commutative, but doesn't have an identity (if $|R| > 1$).

8. Direct sum

Let R and S be any two rings. Then we define the *direct sum* $R \oplus S$ as follows. As a set, $R \oplus S$ is just the cartesian product $R \times S$. The operations are given by the rules

$$(r_1,s_1)+(r_2,s_2) = (r_1+r_2, s_1+s_2), \qquad (r_1,s_1)(r_2,s_2) = (r_1r_2, s_1s_2).$$

(Note that in the ordered pair (r_1+r_2, s_1+s_2), the first $+$ denotes addition in R, and the second $+$ is addition in S.)

Proposition 2.4 *If R and S are rings, then $R \oplus S$ is a ring. If R and S are commutative, then so is $R \oplus S$; if R and S have identities, then so does $R \oplus S$; but $R \oplus S$ is not a division ring if both R and S have more than one element.*

The proof is straightforward checking.

9. Modular arithmetic

Let $\mathbb{Z}_n$ denote the set of all congruence classes modulo n, where n is a positive integer. We saw in the first chapter that there are n congruence classes; so $\mathbb{Z}_n$ is a set with n elements:

$$\mathbb{Z}_n = \{[0]_n, [1]_n, \ldots, [n-1]_n\}.$$

Define addition and multiplication on $\mathbb{Z}_n$ by the rules

$$[a]_n + [b]_n = [a+b]_n, \qquad [a]_n[b]_n = [ab]_n.$$

There is an important job to do here: we have to show that these definitions don't depend on our choice of representatives of the equivalence classes.

Proposition 2.5 *For any positive integer n, $\mathbb{Z}_n$ is a commutative ring with identity. It is a field if and only if n is a prime number.*

Here, for example, are the addition and multiplication tables of the ring $\mathbb{Z}_5$. We simplify the notation by writing x instead of $[x]_5$.

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\cdot$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Note, for example, that $2^{-1} = 3$ in this ring.

10. Rings of functions

The sum and product of continuous real functions are continuous. So there is a ring $C(\mathbb{R})$ of coninuous functions from $\mathbb{R}$ to $\mathbb{R}$, with

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

There are several related rings, such as $C^1(\mathbb{R})$ (the ring of differentiable functions), $C_0(\mathbb{R})$ (the ring of continuous functions satisfying $f(x) \to 0$ as $x \to \pm\infty$), and $C([a,b])$ (the ring of continuous functions on the interval $[a,b]$. All these rings are commutative, and all except $C_0(\mathbb{R})$ have an identity (the constant function with value 1).

These rings are the subject-matter of Functional Analysis.

2.1.3 Properties of rings

We have some business deferred from earlier to deal with. After that, we prove some basic properties of rings, starting from the axioms.

Uniqueness of zero element

The zero element of a ring is unique. For suppose that there are two zero elements, say z_1 and z_2. (This means that $a+z_1 = z_1+a = a$ for all a and also $a+z_2 = z_2+a = a$ for all a.) Then

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

Exercise: Show that the identity element of a ring, if it exists, is unique.

Uniqueness of additive inverse

The additive inverse of an element a is unique. For suppose that b and c are both additive inverses of a. (This means that $a+b = b+a = 0$ and $a+c = c+a = 0$ – we know now that there is a unique zero element, and we call it 0.) Then

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

where we use the associative law in the third step.

Exercise: Show that the multiplicative inverse of an element of a ring, if it exists, is unique.

Adding more than two elements

The associative law tells us that if we have to add three elements, then the two possible ways of doing it, namely $(a+b)+c$ and $a+(b+c)$, give us the same result. For more than three elements, there are many different ways of adding them: we have to put in brackets so that the sum can be worked out by adding two elements at a time. For example, there are five ways of adding four elements:

$$((a+b)+c)+d,\ (a+(b+c))+d,\ (a+b)+(c+d),\ a+((b+c)+d),\ a+(b+(c+d)).$$

These are all equal. For the associative law $(a+b)+c = a+(b+c)$ shows that the first and second are equal, while the associative law for b,c,d shows that the fourth and fifth are equal. Also, putting $x = a+b$, we have

$$((a+b)+c)+d = (x+c)+d = x+(c+d) = (a+b)+(c+d),$$

so the first and third are equal; and similarly the third and fifth are equal.

In general we have the following. The proof works for any associative binary operation.

Proposition 2.6 *Let $*$ be an associative binary operation on a set A, and $a_1,\ldots,a_n \in A$. Then the result of evaluating $a_1 * a_2 * \cdots * a_n$, by adding brackets in any way to make the expression well-defined, is the same, independent of bracketing.*

Proof The proof is by induction on the number of terms. For $n = 2$ there is nothing to prove; for $n = 3$, the statement is just the associative law; and for $n = 4$, we showed it above. Suppose that the result is true for fewer than n terms. Suppose now that we have two different bracketings of the expression $a_1 * a_2 * \cdots * a_n$. The first will have the form $(a_1 * \cdots * a_i) * (a_{i+1} * \cdots * a_n)$, with the terms inside the two sets of brackets themselves bracketed in some way. By induction, the result is independent of the bracketing of $a_1,\ldots,a_i$ and of $a_{i+1},\ldots,a_n$. Similarly, the second expression will have the form $(a_1 * \cdots * a_j) * (a_{j+1} * \cdots * a_n)$, and is independent of the bracketing of $a_1,\ldots,a_j$ and of $a_{j+1},\ldots,a_n$.

Case 1 : $i = j$. Then the two expressions are obviously equal.

Case 2 : $i \neq j$; suppose, without loss, that $i < j$. Then the first expression can be written as

$$(a_1 * \cdots * a_i) * ((a_{i+1} * \cdots * a_j) * (a_{j+1} * \cdots * a_n)),$$

and the second as

$$((a_1 * \cdots * a_i) * (a_{i+1} * \cdots * a_j)) * (a_{j+1} * \cdots * a_n),$$

where each expression is independent of any further bracketing. By the associative law, these two expressions are equal: they are $x * (y * z)$ and $(x * y) * z$, where $x = a_1 * \cdots * a_i$, $y = a_{i+1} * \cdots * a_j$, and $z = a_{j+1} * \cdots * a_n$.

Note that this result applies to both addition and multiplication in a ring.

As usual, we denote $a_1 + a_2 + \cdots + a_n$ by $\sum_{i=1}^{n} a_i$.

Cancellation laws

Proposition 2.7 *In a ring R, if $a+x = b+x$, then $a = b$. Similarly, if $x+a = x+b$, then $a = b$.*

Proof Suppose that $a+x = b+x$, and let $y = -x$. Then

$$a = a+0 = a+(x+y) = (a+x)+y = (b+x)+y = b+(x+y) = b+0 = b.$$

The other law is proved similarly, or by using the commutativity of addition.

These facts are the *cancellation laws*.

A property of zero

One familiar property of the integers is that $0a = 0$ for any integer a. We don't have to include this as an axiom, since it follows from the other axioms. Here is the proof. We have $0+0 = 0$, so $0a+0 = 0a = (0+0)a = 0a+0a$, by the distributive law; so the cancellation law gives $0 = 0a$. Similarly $a0 = 0$.

It follows that if R has an identity 1, and $|R| > 1$, then $1 \neq 0$. For choose any element $a \neq 0$; then $1a = a$ and $0a = 0$. It also explains why we have to exclude 0 in condition (M3): 0 cannot have a multiplicative inverse.

Commutativity of addition

It turns out that, in a ring with identity, it is not necessary to assume that addition is commutative: axiom (A4) follows from the other ring axioms together with (M2).

For suppose that (A0)–(A3), (M0)–(M2) and (D) all hold. We have to show that $a+b = b+a$. Consider the expression $(1+1)(a+b)$. We can expand this in two different ways by the two distributive laws:

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

Hence $a+b+a+b = a+a+b+b$, and using the two cancellation laws we conclude that $b+a = a+b$.

This argument depends on the existence of a multiplicative identity. If we take a structure with an operation $+$ satisfying (A0)–(A3) (we'll see later that such a structure is known as a *group*), and apply the "zero ring" construction to it (that is, $ab = 0$ for all a, b), we obtain a structure satisfying all the ring axioms except (A4).

Boolean rings

We saw that a *Boolean ring* is a ring R in which $xx = x$ for all $x \in R$.

Proposition 2.8 *A Boolean ring is commutative and satisfies $x+x = 0$ for all $x \in R$.*

Proof We have $(x+y)(x+y) = x+y$. Expanding the left using the distributive laws, we find that

$$xx+xy+yx+yy = x+y.$$

Now $xx = x$ and $yy = y$. So we can apply the cancellation laws to get

$$xy+yx = 0.$$

In particular, putting $y = x$ in this equation, we have $xx+xx = 0$, or $x+x = 0$, one of the things we had to prove.

Taking this equation and putting xy in place of x, we have

$$xy+xy = 0 = xy+yx,$$

and then the cancellation law gives us $xy = yx$, as required.

We saw that the power set of any set, with the operations of symmetric difference and intersection, is a Boolean ring. Another example is the ring $\mathbb{Z}_2$ (the integers mod 2).

2.1.4 Matrix rings

In view of Proposition 2.6, the definition of the product of two $n \times n$ matrices now makes sense: $AB = D$, where

$$D_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}.$$

So we are in the position to prove Proposition 2.1.

A complete proof of this proposition involves verifying all the ring axioms. The arguments are somewhat repetitive; I will give proofs of two of the axioms.

Axiom (A2): Let 0 be the zero element of the ring R, and let O be the zero matrix in $M_n(R)$, satisfying $O_{ij} = 0$ for all i, j. Then O is the zero element of $M_n(R)$: for, given any matrix A,

$$(O+A)_{ij} = O_{ij} + A_{ij} = 0 + A_{ij} = A_{ij}, \qquad (A+O)_{ij} = A_{ij} + O_{ij} = A_{ij} + 0 = A_{ij},$$

using the properties of $0 \in R$. So $O + A = A + O = A$.

Axiom (D): the (i, j) entry of $A(B+C)$ is

$$\sum_{k=1}^{n} A_{ik}(B+C)_{kj} = \sum_{k=1}^{n} A_{ik}B_{kj} + A_{ik}C_{kj},$$

by the distributive law in R; and the (i, j) entry of $AB + AC$ is

$$\sum_{k=1}^{n} A_{ik}B_{kj} + \sum_{k=1}^{n} A_{ik}C_{kj}.$$

Why are these two expressions the same? Let us consider the case $n = 2$. The first expression is

$$A_{i1}B_{1j} + A_{i1}C_{1j} + A_{i2}B_{2j} + A_{i2}C_{2j},$$

while the second is

$$A_{i1}B_{1j} + A_{i2}B_{2j} + A_{i1}C_{1j} + A_{i2}C_{2j}.$$

(By Proposition 2.6, the bracketing is not significant.) Now the commutative law for addition allows us to swap the second and third terms of the sum; so the two expressions are equal. Hence $A(B+C) = AB + AC$ for any matrices A, B, C. For $n > 2$, things are similar, but the rearrangement required is a bit more complicated.

The proof of the other distributive law is similar.

Observe what happens in this proof: we use properties of the ring R to deduce properties of $M_n(R)$. To prove the distributive law for $M_n(R)$, we needed the distributive law and the associative and commutative laws for addition in R. Similar things happen for the other axioms.

2.1.5 Polynomial rings

What exactly is a polynomial? We deferred this question before, but now is the time to face it.

A polynomial $\sum a_i x^i$ is completely determined by the sequence of its coefficients $a_0, a_1, \ldots$. These have the property that only a finite number of terms in the sequence are non-zero, but we cannot say in advance how many. So we make the following definition:

A *polynomial* over a ring R is an infinite sequence

$$(a_i)_{i \geq 0} = (a_0, a_1, \ldots)$$

of elements of R, having the property that only finitely many terms are non-zero; that is, there exists an n such that $a_i = 0$ for all $i > n$. If a_n is the last non-zero term, we say that the *degree* of the polynomial is n. (Note that, according to this definition, the all-zero sequence does not have a degree.)

Now the rules for addition and multiplication are

$$(a_i) + (b_i) = (c_i) \quad \text{where} \quad c_i = a_i + b_i,$$
$$(a_i)(b_i) = (d_i) \quad \text{where} \quad d_i = \sum_{j=0}^{i} a_j b_{i-j}.$$

Again, the sum in the definition of multiplication is justified by Proposition 2.6. We think of the polynomial $(a_i)_{i \geq 0}$ of degree n as what we usually write as $\sum_{i=0}^{n} a_i x^i$; the rules we gave agree with the usual ones.

Now we can prove Proposition 2.2, asserting that the set of polynomials over a ring R is a ring. As for matrices, we have to check all the axioms, which involves a certain amount of tedium. The zero polynomial required by (A2) is the all-zero sequence. Here is a proof of (M1). You will see that it involves careful work with dummy subscripts!

We have to prove the associative law for multiplication. So suppose that $f = (a_i)$, $g = (b_i)$ and $h = (c_i)$. Then the ith term of fg is $\sum_{j=0}^{i} a_j b_{i-j}$, and so the ith term of $(fg)h$ is

$$\sum_{k=0}^{i} \left(\sum_{j=0}^{k} a_j b_{k-j} \right) c_{i-k}.$$

Similarly the ith term of $f(gh)$ is

$$\sum_{s=0}^{i} a_s \left(\sum_{t=0}^{i-s} b_t c_{i-s-t} \right).$$

Each term on both sides has the form $a_p b_q c_r$, where $p, q, r \geq 0$ and $p + q + r = i$. (In the first expression, $p = j$, $q = k - j$, $r = i - k$; in the second, $p = s$, $q = t$,

$r = i - s - t$.) So the two expressions contain the same terms in a different order. By the associative and commutative laws for addition, they are equal.

2.2 Subrings

2.2.1 Definition and test

Suppose that we are given a set S with operations of addition and multiplication, and we are asked to prove that it is a ring. In general, we have to check all the axioms. But there is a situation in which things are much simpler: this is when S is a subset of a set R which we already know to be a ring, and the addition and multiplication in S are just the restrictions of the operations in R (that is, to add two elements of S, we regard them as elements of R and use the addition in R).

Definition Let R be a ring. A *subring* of R is a subset S of R which is a ring in its own right with respect to the restrictions of the operations in R.

What do we have to do to show that S is a subring?

- The associative law (A1) holds in S. For, if $a, b, c \in S$, then we have $a, b, c \in R$ (since $S \subseteq R$), and so

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

  since R satisfies (A1) (as we are given that it is a ring).

- Exactly the same argument shows that the commutative law for addition (A4), the associative law for multiplication (M1), and the distributive laws (D), all hold in S.

- This leaves only (A0), (A2), (A3) and (M0) to check.

Even here we can make a simplification, if $S \neq \emptyset$. For suppose that (A0) and (A3) hold in S. Given $a \in S$, the additive inverse $-a$ belongs to S (since we are assuming (A3)), and so $0 = a + (-a)$ belongs to S (since we are assuming (A0)). Thus (A2) follows from (A0) and (A3).

We state this as a theorem:

Theorem 2.9 (First Subring Test) *Let R be a ring, and let S be a non-empty subset of R. Then S is a subring of R if the following condition holds:*

for all $a, b \in S$, we have $a+b, ab, -a \in S$.

Example We show that the set S of even integers is a ring. Clearly it is a non-empty subset of the ring $\mathbb{Z}$ of integers. Now, if $a, b \in S$, say $a = 2c$ and $b = 2d$, we have

$$a+b = 2(c+d) \in S, \qquad ab = 2(2cd) \in S, \qquad -a = 2(-c) \in S,$$

and so S is a subring of $\mathbb{Z}$, and hence is a ring.

The theorem gives us three things to check. But we can reduce the number from three to two. We use $a-b$ as shorthand for $a+(-b)$. In the next proof we need to know that $-(-b) = b$. This holds for the following reason. We have, by (A3),

$$b+(-b) = (-b)+b = 0,$$

so that b is an additive inverse of $-b$. Also, of course, $-(-b)$ is an additive inverse of $-b$. By the uniqueness of additive inverse, $-(-b) = b$, as required. In particular, $a-(-b) = a+(-(-b)) = a+b$.

Theorem 2.10 (Second Subring Test) *Let R be a ring, and let S be a non-empty subset of R. Then S is a subring of R if the following condition holds:*

for all $a, b \in S$, we have $a-b, ab \in S$.

Proof Let S satisfy this condition: that is, S is closed under subtraction and multiplication. We have to verify that it satisfies the conditions of the First Subring Test. Choose any element $a \in S$ (this is possible since S is non-empty). Then the hypothesis of the theorem shows that $0 = a - a \in S$. Applying the hypothesis again shows that $-a = 0 - a \in S$. Finally, if $a, b \in S$, then $-b \in S$ (by what has just been proved), and so $a+b = a-(-b) \in S$. So we are done.

2.2.2 Cosets

Suppose that S is a subring of R. We now define a partition of R, one of whose parts is S. Remember that, by the Equivalence Relation Theorem, in order to specify a partition of R, we must give an equivalence relation on R.

Let $\equiv_S$ be the relation on R defined by the rule

$$a \equiv_S b \quad \text{if and only if} \quad b - a \in S.$$

We claim that $\equiv_S$ is an equivalence relation.

Reflexive: for any $a \in R$, $a - a = 0 \in S$, so $a \equiv_S a$.

Symmetric: take $a, b \in R$ with $a \equiv_S b$, so that $b - a \in S$. Then $a - b = -(b-a) \in S$, so $b \equiv_S a$.

Transitive: take $a,b,c \in R$ with $a \equiv_S b$ and $b \equiv_S c$. Then $b-a, c-b \in S$. So $c-a = (c-b)+(b-a) \in S$, so $a \equiv_S c$.

So $\equiv_S$ is an equivalence relation. Its equivalence classes are called the *cosets* of S in R.

Example Let n be a positive intger. Let $R = \mathbb{Z}$ and $S = n\mathbb{Z}$, the set of all multiples of n. Then S is a subring of R. (By the Second Subring Test, if $a,b \in S$, say $a = nc$ and $b = nd$, then $a-b = n(c-d) \in S$ and $ab = n(ncd) \in S$.) In this case, the relation $\equiv_S$ is just congruence mod n, since $a \equiv_S b$ if and only if $b-a$ is a multiple of n. The cosets of S are thus precisely the congruence classes mod n.

An element of a coset is called a *coset representative*. As we saw in the first chapter, it is a general property of equivalence relations that any element can be used as the coset representative: if b is in the same equivalence class as a, then a and b define the same equivalence classes. We now give a description of cosets.

If S is a subset of R, and $a \in R$, we define $S+a$ to be the set

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

consisting of all elements that we can get by adding a to an element of S.

Proposition 2.11 *Let S be a subring of R, and $a \in R$. Then the coset of R containing a is $S+a$.*

Proof Let $[a]$ denote the coset containing a, that is,

$$[a] = \{b \in R : a \equiv_S b\} = \{b \in R : b-a \in S\}.$$

We have to show that $[a] = S+a$.

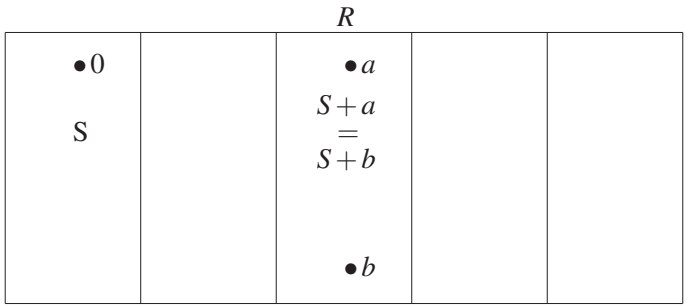
First take $b \in [a]$, so that $b-a \in S$. Let $s = b-a$. Then $b = s+a \in S+a$.

In the other direction, take $b \in S+a$, so that $b = s+a$ for some $s \in S$. Then $b-a = (s+a)-a = s \in S$, so $b \equiv_S a$, that is, $b \in [a]$.

So $[a] = S+a$, as required.

Any element of a coset can be used as its representative. That is, if $b \in S+a$, then $S+a = S+b$.

Here is a picture.



Note that $S+0 = S$, so the subring S is a coset of itself, namely the coset containing 0.

In particular, the congruence class $[a]_n$ in $\mathbb{Z}$ is the coset $n\mathbb{Z}+a$, consisting of all elements obtained by adding a multiple of n to a. So the ring $\mathbb{Z}$ is partitioned into n cosets of $n\mathbb{Z}$.

2.3 Homomorphisms and quotient rings

2.3.1 Isomorphism

Here are the addition and multiplication tables of a ring with two elements, which for now I will call o and i.

$+$	o	i
o	o	i
i	i	o

$\cdot$	o	i
o	o	o
i	o	i

You may recognise this ring in various guises: it is the Boolean ring $\mathscr{P}(X)$, where $X = \{x\}$ is a set with just one element x; we have $o = \emptyset$ and $i = \{x\}$. Alternatively it is the ring of integers mod 2, with $o = [0]_2$ and $i = [1]_2$.

The fact that these two rings have the same addition and multiplication tables shows that, from an algebraic point of view, we cannot distinguish between them.

We formalise this as follows. Let R_1 and R_2 be rings. Let $\theta : R_1 \to R_2$ be a function which is one-to-one and onto, that is, a bijection between R_1 and R_2. Now we denote the result of applying the function θ to an element $r \in R_1$ by $r\theta$ or $(r)\theta$ rather than by $\theta(r)$; that is, we write the function on the right of its argument.

Now we say that θ is an *isomorphism* from R_1 to R_2 if it is a bijection which satisfies

$$(r_1+r_2)\theta = r_1\theta + r_2\theta, \qquad (r_1r_2)\theta = (r_1\theta)(r_2\theta). \tag{2.1}$$

This means that we "match up" elements in R_1 with elements in R_2 so that addition and multiplication work in the same way in both rings.

Example To return to our earlier example, let $R_1 = \mathscr{P}(\{x\})$ and let R_2 be the ring of integers mod 2, and define a function $\theta : R_1 \to R_2$ by

$$\emptyset\theta = [0]_2, \qquad \{x\}\theta = [1]_2.$$

Then θ is an isomorphism.

We say that the rings R_1 and R_2 are "isomorphic" if there is an isomorphism from R_1 to R_2. The word "isomorphic" means, roughly speaking, "the same shape": if two rings are isomorphic then they can be regarded as identical from the point of view of Ring Theory, even if their actual elements are quite different (as in our example). We could say that Ring Theory is the study of properties of rings which are the same in isomorphic rings.

So, for example, if R_1 and R_2 are isomorphic then:

- If R_1 is commutative, then so is R_2, and vice versa; and the same holds for the property of being a ring with identity, a division ring, a Boolean ring, a zero ring, etc.
- However, the property of being a ring of matrices, or a ring of polynomials, etc., are not necessarily shared by isomorphic rings.

We use the notation $R_1 \cong R_2$ to mean "R_1 is isomorphic to R_2". Remember that isomorphism is a relation between two rings. If you are given two rings R_1 and R_2 and asked whether they are isomorphic, **do not** say "R_1 is isomorphic but R_2 is not".

2.3.2 Homomorphisms

An isomorphism is a function between rings with two properties: it is a bijection (one-to-one and onto), and it preserves addition and multiplication (as expressed by equation (2.1)). A function which preserves addition and multiplication but is not necessarily a bijection is called a homomorphism. Thus, a *homomorphism* from R_1 to R_2 is a function $\theta : R_1 \to R_2$ satisfying

$$(r_1 + r_2)\theta = r_1\theta + r_2\theta, \qquad (r_1 r_2)\theta = (r_1\theta)(r_2\theta).$$

You should get used to these two long words, and two others. A function $\theta : R_1 \to R_2$ is

- a *homomorphism* if it satisfies (2.1); (homo=similar)
- a *monomorphism* if it satisfies (2.1) and is one-to-one; (mono=one)
- an *epimorphism* if it satisfies (2.1) and is onto; (epi=onto)

- an *isomorphism* if it satisfies (2.1) and is one-to-one and onto (iso=equal)

For example, the function from the ring $\mathbb{Z}$ to the ring of integers mod 2, which takes the integer n to its congruence class $[n]_2$ mod 2, is a homomorphism. Basically this says that, if we only care about the parity of an integer, its congruence mod 2, then the addition and multiplication tables are

$+$	even	odd
even	even	odd
odd	odd	even

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

and this ring is the same as the one at the start of this section.

Let $\theta : R_1 \to R_2$ be a homomorphism. The *image* of θ is, as usual, the set

$$\mathrm{Im}(\theta) = \{s \in R_2 : s = r\theta \text{ for some } r \in R_1\}.$$

We define the *kernel* of θ to be the set

$$\mathrm{Ker}(\theta) = \{r \in R_1 : r\theta = 0\},$$

the set of elements of R_1 which are mapped to the zero element of R_2 by θ. You will have seen a definition very similar to this in Linear Algebra.

The image and kernel of a homomorphism have an extra property. This is not the final version of this theorem: we will strengthen it in two ways in the next two sections. First, a lemma:

Lemma 2.12 *Let $\theta : R_1 \to R_2$ be a homomorphism. Then*

(a) $0\theta = 0$*;*

(b) $(-a)\theta = -(a\theta)$ *for all* $a \in R_1$*;*

(c) $(a-b)\theta = a\theta - b\theta$ *for all* $a, b \in R_1$*.*

Proof We have

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

and the cancellation law gives $0\theta = 0$.

Then

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

so $(-a)\theta$ is the additive inverse of $a\theta$, that is, $(-1)\theta = -(a\theta)$.

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

Proposition 2.13 *Let $\theta : R_1 \to R_2$ be a homomorphism. Then*

(a) $\mathrm{Im}(\theta)$ *is a subring of* R_2*;*

(b) $\mathrm{Ker}(\theta)$ *is a subring of* R_1.

Proof We use the Second Subring Test.

(a) First notice that $\mathrm{Im}(\theta) \neq \emptyset$, since $\mathrm{Im}(\theta)$ contains 0, by the Lemma.
Take $a, b \in \mathrm{Im}(\theta)$, say, $a = x\theta$ and $b = y\theta$. Then $-b = (-y)\theta$, so

$$a - b = x\theta + (-y)\theta = (x-y)\theta \in \mathrm{Im}(\theta).$$

Also $ab = (x\theta)(y\theta) = (xy)\theta \in \mathrm{Im}(\theta)$. So $\mathrm{Im}(\theta)$ is a subring of R_2.

(b) First notice that $\mathrm{Ker}(\theta) \neq \emptyset$, since $\mathrm{Ker}(\theta)$ contains 0, by the Lemma.
Take $a, b \in \mathrm{Ker}(\theta)$, so that $a\theta = b\theta = 0$. Then

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

so $\mathrm{Ker}(\theta)$ is a subring.

2.3.3 Ideals

An ideal in a ring is a special kind of subring.

Let S be a subring of R. We say that S is an *ideal* if, for any $a \in S$ and $r \in R$, we have $ar \in S$ and $ra \in S$.

For example, let $R = \mathbb{Z}$ and $S = n\mathbb{Z}$ for some positive integer n. We know that S is a subring of R. Choose $a \in S$, say $a = nc$ for some $c \in \mathbb{Z}$. Then $ar = ra = n(cr) \in S$. So S is an ideal.

Any ring R has two trivial ideals: the whole ring R is an ideal; and the set $\{0\}$ consisting only of the zero element is an ideal.

There is an ideal test similar to the subring tests. We give just one form.

Theorem 2.14 (Ideal Test) *Let R be a ring, and S a non-empty subset of R. Then S is an ideal if the following conditions hold:*

(a) for all $a, b \in S$, we have $a - b \in S$;

(b) for all $a \in S$ and $r \in R$, we have $ar, ra \in S$.

Proof Take $a, b \in S$. Then $ab \in S$ (this is a special case of (b), with $r = b$). So by the Second Subring Test, S is a subring. Then by (b), it is an ideal.

Now we can strengthen the statement that the kernel of a homomorphism is a subring.

Proposition 2.15 *Let $\theta : R_1 \to R_2$ be a homomorphism. Then* $\mathrm{Ker}(\theta)$ *is an ideal in R_1.*

Proof We already know that it is a subring, so we only have to check the last part of the definition. So take $a \in \mathrm{Ker}(\theta)$ (so that $a\theta = 0$), and $r \in R_1$. Then

$$(ar)\theta = (a\theta)(r\theta) = 0(r\theta) = 0,$$

and similarly $(ra)\theta = 0$. So $ar, ra \in \mathrm{Ker}(\theta)$.

We will see in the next section that it goes the other way too: every ideal is the kernel of a homomorphism. So "ideals" are the same thing as "kernels of homomorphisms".

2.3.4 Quotient rings

Let I be an ideal of a ring R. We will define a ring, which we call the *quotient ring* or *factor ring*, of R by I, and denote by R/I.

The elements of R/I are the cosets of I in R. Thus each element of R/I is a set of elements (an equivalence class) of R. Remember that each coset can be written as $I + a$ for some $a \in R$. Now we have to define addition and multiplication. We do this by the rules

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

There is one important job that we have to do to prove that this is a good definition. Remember that any element of a coset can be used as a representative. So you might use the representatives a and b, while I use the representatives a' and b' for the same cosets. We need to show that the definitions don't depend on these choices; that is, we have to show that

$$I+a = I+a' \text{ and } I+b = I+b' \text{ imply } I+(a+b) = I+(a'+b') \text{ and } I+ab = I+a'b'.$$

So suppose that $I + a = I + a'$ and $I + b = I + b'$. Then $a' \in I + a$, so $a' = s + a$ for some $s \in I$. Similarly, $b' = t + b$ for some $t \in I$. Now

$$\begin{aligned} a'+b' = (s+a)+(t+b) &= (s+t)+(a+b) \in I+(a+b), \\ a'b' = (s+a)(t+b) &= st+sb+ta+ab \in I+ab, \end{aligned}$$

by using the associative and commutative laws for addition and the distributive laws. So the result is proved, once we justify the last step by showing that $s+t \in I$ and $st+sb+at \in I$. Remember that $s,t \in I$, so that $s+t \in I$ (as I is a subring); also $st \in I$ (since I is a subring) and $sb \in I$ and $at \in I$ (since I is an ideal), so the sum of these three expressions is in I.

Proposition 2.16 *If I is an ideal of the ring R, then the set R/I, with operations of addition and multiplication defined as above, is a ring, and the map $\theta : R \to R/I$ defined by $r\theta = I+r$ is a homomorphism whose kernel is I.*

Proof We have well-defined operations of addition and multiplication, so (A0) and (M0) hold. The proofs of the other axioms are all very similar. Here is a proof of the first distributive law. Take three elements of R/I (that is, three cosets!), say $I+a, I+b, I+c$. Then

$$\begin{aligned}((I+a)+(I+b))(I+c) &= (I+(a+b))(I+c)\\ &= I+(a+b)c\\ &= I+(ac+bc)\\ &= (I+ac)+(I+bc)\\ &= (I+a)(I+c)+(I+b)(I+c).\end{aligned}$$

Here we use the distributive law in R to get from the second line to the third, while the other steps just use the definitions of addition and multiplication in R/I.

Next we show that θ is a homomorphism. This is true by definition:

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

Finally we calculate $\mathrm{Ker}(\theta)$. There is one important thing to note. The zero element of R/I is the coset $I+0$. This is just the ideal I itself! So

$$\mathrm{Ker}(\theta) = \{a \in R : a\theta = 0\} = \{a \in R : I+a = I\} = I,$$

since $I+a = I$ means that a is a representative for the coset I, that is, $a \in I$.

The map θ in this result is called the *natural homomorphism* from R to R/I. We see that, if I is any ideal of R, then I is the kernel of the natural homomorphism from R to R/I.

2.3.5 The Isomorphism Theorems

The Isomorphism Theorems are a number of results which look more closely at a homomorphism. The first one makes more precise the results we saw earlier about the image and kernel of a homomorphism.

Theorem 2.17 (First Isomorphism Theorem) *Let R_1 and R_2 be rings, and let $\theta : R_1 \to R_2$ be a homomorphism. Then*

(a) $\mathrm{Im}(\theta)$ *is a subring of R_2;*

(b) $\mathrm{Ker}(\theta)$ *is an ideal of R_1;*

(c) $R_1/\mathrm{Ker}(\theta) \cong \mathrm{Im}(\theta)$.

Proof We already proved the first two parts of this theorem, in Propositions 2.13 and 2.15. We have to prove (c). Remember that this means that the rings $R_1/\mathrm{Ker}(\theta)$ (the quotient ring, which is defined because $\mathrm{Ker}(\theta)$ is an ideal in R_1) and $\mathrm{Im}(\theta)$ (a subring of R_2) are isomorphic. We have to construct a map ϕ between these two rings which is one-to-one and onto, and is a homomorphism.

Put $I = \mathrm{Ker}(\theta)$, and define ϕ by the rule

$$(I+r)\phi = r\theta$$

for $r \in R_1$. On the face of it, this might depend on the choice of the coset representative r. So first we have to prove that, if $I+r = I+r'$, then $r\theta = r'\theta$. We have

$$\begin{aligned}I+r = I+r' &\Rightarrow r' = s+r \text{ for some } s \in I = \mathrm{Ker}(\theta)\\ &\Rightarrow r'\theta = s\theta + r\theta = 0 + r\theta = r\theta,\end{aligned}$$

as required. So indeed ϕ is well defined.

In fact this argument also reverses. If $r\theta = r'\theta$, then $(r'-r)\theta = r'\theta - r\theta = 0$, so $r'-r \in \mathrm{Ker}(\theta)$. This means, by definition, that r and r' lie in the same coset of $\mathrm{Ker}(\theta) = I$, so that $I+r = I+r'$. This shows that ϕ is one-to-one.

To show that ϕ is onto, take $s \in \mathrm{Im}(\theta)$. Then $s = r\theta$ for some $r \in R$, and we have $s = r\theta = (I+r)\phi$. So $\mathrm{Im}(\phi) = \mathrm{Im}(\theta)$ as required.

Finally,

$$\begin{aligned}((I+r_1)+(I+r_2))\phi = (r_1+r_2)\theta &= (r_1\theta)+(r_2\theta) = (I+r_1)\phi+(I+r_2)\phi,\\ ((I+r_1)(I+r_2))\phi = (r_1r_2)\theta &= (r_1\theta)(r_2\theta) = (I+r_1)\phi(I+r_2)\phi,\end{aligned}$$

so ϕ is a homomorphism, and hence an isomorphism, as required.

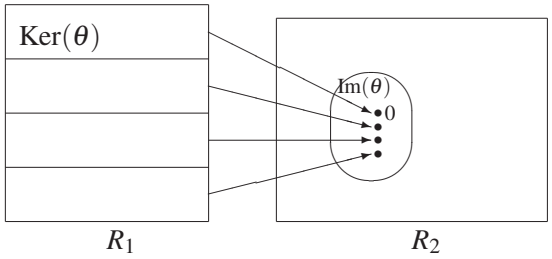


Figure 2.1: A homomorphism

We illustrate this theorem with a picture.

In the picture, the parts into which R_1 is divided are the cosets of the ideal $\ker(\theta)$ (the set $\mathrm{Ker}(\theta)$ itself has been taken to be the top part of the partition). The oval region inside R_2 is the subring $\mathrm{Im}(\theta)$. Each coset of $\mathrm{Ker}(\theta)$ maps to a single element of $\mathrm{Im}(\theta)$.

The second Isomorphism Theorem is sometimes called the "Correspondence Theorem", since it says that subrings of R/I correspond in a one-to-one manner with subrings of R containing I.

Theorem 2.18 (Second Isomorphism Theorem) *Let I be an ideal of the ring R. Then there is a one-to-one correspondence between the subrings of R/I and the subrings of R containing I, given as follows: to a subring S of R containing I corresponds the subring S/I of R/I. Under this correspondence, ideals of R/I correspond to ideals of R containing I; and, if J is an ideal of R containing I, then*

$$(R/I)/(J/I) \cong R/J.$$

Proof If S is a subring of R containing I, then I is an ideal of S. (For applying the ideal test inside S means we have to check that I is closed under subtraction and under multiplication by elements of S; these are just some of the checks that would be required to show that it is an ideal of R. Now if $s \in S$, then the entire coset $I+s$ lies in S, since S is closed under addition. So S/I is well-defined: it consists of all the cosets of I which are contained in S. Clearly it is a subring of R/I. Thus, we have a mapping from subrings of R containing I to subrings of R/I.

In the other direction, let T be a subring of R/I. This means that T is a set of cosets of I which form a ring. Let S be the *union* of all the cosets in T. We will show that S is a subring of R. It obviously contains I (since I is the zero coset) and $S/I = T$ follows.

Take $a,b \in S$. Then $I+a, I+b \in T$. Since T is a subring, we have $(I+a)-(I+b) = I+(a-b) \in T$ and $(I+a)(I+b) = I+ab \in T$, so $a-b \in S$ and $ab \in S$. By the Second Subring Test, S is a subring.

Next we show that ideals correspond to ideals. Let J be an ideal of R containing I. Then J/I is a subring of R/I, and we have to show that it is an ideal. Take $I+a \in J/I$ and $I+r \in R/I$. Then $a \in J$ and $r \in R$, so $ar, ra \in J$, whence $(I+a)(I+r), (I+r)(I+a) \in J/I$. Thus J/I is an ideal of R/I. The converse is similar.

I will not give the proof that $(R/I)/(J/I) \cong R/J$: this will not be used in the course.

The Third Isomorphism Theorem needs a little more notation. Let A and B be two subsets of a ring R. Then we define $A+B$ to consist of all sums of an element of A and an element of B:

$$A+B = \{a+b : a \in A, b \in B\}.$$

Theorem 2.19 (Third Isomorphism Theorem) *Let R be a ring, S a subring of R, and I an ideal of R. Then*

(a) $S+I$ is a subring of R containing I;

(b) $S \cap I$ is an ideal of S;

(c) $S/(S \cap I) \cong (S+I)/I$.

Proof We could prove the three parts in order, but it is actually easier to start at the end! Remember the natural homomorphism θ from R to R/I with kernel θ. What happens when we restrict θ to S, that is, we only put elements of S into the function θ? Let ϕ denote this restriction. Then ϕ maps S to R/I. We find its image and kernel, and apply the First Isomorphism Theorem to them.

(a) The image of ϕ consists of all cosets $I+s$ containing a coset representative in S. The union of all these cosets is $I+S$, so the image of ϕ is $(I+S)/I$. This is a subring of R/I (since it is the image of a homomorphism). By the Correspondence Theorem, $S+I$ is a subring of R containing I.

(b) The kernel of ϕ consists of all elements of S mapped to zero by ϕ, that is, all elements $s \in S$ such that $s \in \mathrm{Ker}(\theta) = I$. Thus, $\mathrm{Ker}(\phi) = S \cap I$, and so $S \cap I$ is an ideal of S.

(c) Now the first isomorphism theorem shows that

$$S/(I+S) \cong \mathrm{Im}(\phi) = (I+S)/I,$$

and we are done.

2.4 Factorisation

One of the most important properties of the integers is that any number can be factorised into prime factors in a unique way. But we have to be a bit careful. It would be silly to try to factorise 0 or 1; and the factorisation is not quite unique, since $(-2)\cdot(-3) = 2\cdot 3$, for example. Once we have the definitions straight, we will see that "unique factorisation" holds in a large class of rings.

2.4.1 Zero divisors and units

In this section, we will assume that our rings are always commutative.

Let R be a ring. We know that $0a = 0$ holds for all $a \in R$. It is also possible for the product of two non-zero elements of R to be zero. We say that a is a *zero-divisor* if

- $a \neq 0$, and
- there exists $b \in R$, with $b \neq 0$, such that $ab = 0$.

In other words, if the product of two non-zero elements is zero, then we call each of them a zero-divisor.

The ring $\mathbb{Z}$ has no zero-divisors, since if a and b are non-zero integers then obviously $ab \neq 0$. Also, a field has no zero divisors. For suppose that R is a field, and let a be a zero-divisor. Thus, $a \neq 0$, and there exists $b \neq 0$ such that $ab = 0$. Since R is a field, a has a multiplicative inverse a^{-1} satisfying $a^{-1}a = 1$. Then

$$0 = a^{-1}0 = a^{-1}(ab) = (a^{-1}a)b = 1b = b,$$

contradicting our assumption that $b \neq 0$.

In the next example, we use the greatest common divisor function for integers: d is a greatest common divisor of a and b if it divides both of them, and if any other divisor of a and b also divides d. That is, 6 is a greatest common divisor of 12 and 18; but -6 is also a greatest common divisor. We will live with this slight awkwardness for a while, choosing $\gcd(a,b)$ to be the positive rather than the negative value.

Example Let $R = \mathbb{Z}/n\mathbb{Z}$, the ring of integers mod n. Then the element $a \in R$ is a zero-divisor if and only if $1 < \gcd(a,n) < n$.

Proof Suppose that a is a zero-divisor in R. This means that $a \neq 0$ in R (that is, a is not divisible by n, which shows that $\gcd(a,n) < n$), and there exists $b \in R$ with $b \neq 0$ and $ab = 0$. So, regarding a, b, n as integers, we have $n \mid ab$ but n

doesn't divide either a or b. We are trying to prove that $\gcd(a,n) > 1$, so suppose (for a contradiction) that the greatest common divisor is 1. Since n and a are coprime, the fact that n divides ab means that n must divide b, which contradicts our assumption that $b \neq 0$ in R.

Conversely, suppose that $1 < d = \gcd(a,n) < n$. Then $a \neq 0$ as an element of R. Let $a = dx$ and $n = db$. Then n divides $nx = (db)x = (dx)b = ab$, but clearly n doesn't divide y. So, in the ring R, we have $ab = 0$ and $b \neq 0$. Thus a is a zero-divisor.

From now on we make another assumption about our rings: as well as being commutative, they will always have an identity element. We make a definition:

An *integral domain* is a commutative ring with identity which has no zero-divisors.

Example $\mathbb{Z}$ is an integral domain. (This example is the "prototype" of an integral domain, and gives us the name for this class of rings.) Any field is an integral domain. The ring $\mathbb{Z}/n\mathbb{Z}$ is an integral domain if and only if n is a prime number.

The last statement is true because a positive integer n has the property that every smaller positive integer a satisfies $\gcd(a,n) = 1$ if and only if n is prime.

Example If R is an integral domain, then so is the ring $R[x]$ of polynomials over R.

For suppose that f and g are non-zero polynomials, with degrees m and n respectively: that is,

$$f(x) = \sum_{i=0}^{n} a_i x^i, \qquad g(x) = \sum_{i=0}^{m} b_i x^i,$$

where $a_n \neq 0$ and $b_m \neq 0$. The coefficient of x^{m+n} in $f(x)g(x)$ is $a_n b_m \neq 0$ (because R is an integral domain). So $f(x)g(x) \neq 0$.

Let R be a ring with identity element 1; we assume that $1 \neq 0$, Let $a \in R$, with $a \neq 0$. An *inverse* of a is an element $b \in R$ such that $ab = ba = 1$. We say that a us a *unit* if it has an inverse. (We exclude zero because obviously 0 has no inverse: $0b = 0$ for any element b.)

An element a has at most one inverse. For suppose that b and c are inverses of a. Then

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

We write the inverse of the unit a as a^{-1}. Furthermore, a zero-divisor cannot be a unit. For, if $ba = 1$ and $ac = 0$, then

$$0 = b0 = b(ac) = (ba)c = 1c = c.$$

Lemma 2.20 *Let R be a ring with identity. Then*

(a) 1 is a unit;

(b) if u is a unit then so is u^{-1};

(c) if u and v are units then so is uv.

Proof (a) $1 \cdot 1 = 1$.

(b) The equations $uu^{-1} = u^{-1}u = 1$ show that the inverse of u^{-1} is u.

(c) Let u and v be units. We claim that the inverse of uv is $v^{-1}u^{-1}$. (Note the reverse order!) For we have

$$\begin{aligned}(uv)(v^{-1}u^{-1}) &= u(vv^{-1})u^{-1} = u1u^{-1} = uu^{-1} = 1,\\ (v^{-1}u^{-1})(uv) &= v^{-1}(u^{-1}u)v = v^{-1}1v = v^{-1}v = 1.\end{aligned}$$

To help you remember that you have to reverse the order when you find the inverse of a product, this example may help. Suppose that u is the operation of putting on your socks, and v the operation of putting on your shoes, so that uv means "put on your socks and then your shoes". What is the inverse of uv?

Example In the integral domain $\mathbb{Z}$, the only units are $+1$ and -1. For if $ab = 1$, then $a = 1$ or $a = -1$.

Example Consider the ring $\mathbb{Z}/n\mathbb{Z}$, where $n > 1$. We already saw that a is a zero-divisor if and only if $1 < \gcd(a,n) < n$. We claim that a is a unit if and only if $\gcd(a,n) = 1$.

Suppose first that a is a unit, and that $d = \gcd(a,n)$. Then $d \mid a$ and $d \mid n$. Let b be the inverse of a, so that $ab = 1$ in R, which means that $ab \equiv 1 \pmod{n}$, or $ab = xn+1$. But then d divides ab and d divides xn, so d divides 1, whence $d = 1$.

To prove the converse, we use the Euclidean algorithm (more about this shortly), which shows that, given any two integers a and n, there are integers x and y such that $xa + yn = d$, where $d = \gcd(a,n)$. If $d = 1$, then this equation shows that $xa \equiv 1 \pmod{n}$, so that $xa = 1$ in $\mathbb{Z}/n\mathbb{Z}$, so that a is a unit.

This shows that every non-zero element of $\mathbb{Z}/n\mathbb{Z}$ is either a zero-divisor or a unit.

For example, for $n = 12$, we have:

1	unit	$1 \cdot 1 = 1$
2	zero-divisor	$2 \cdot 6 = 0$
3	zero-divisor	$3 \cdot 4 = 0$
4	zero-divisor	$4 \cdot 3 = 0$
5	unit	$5 \cdot 5 = 1$
6	zero-divisor	$6 \cdot 2 = 0$
7	unit	$7 \cdot 7 = 1$
8	zero-divisor	$8 \cdot 3 = 0$
9	zero-divisor	$9 \cdot 4 = 0$
10	zero-divisor	$10 \cdot 6 = 0$
11	unit	$11 \cdot 11 = 1$

We call two elements $a, b \in R$ *associates* if there is a unit $u \in R$ such that $b = ua$. Write $a \sim b$ to mean that a and b are associates. Thus, any unit is an associate of 1, while 0 is associate only to itself.

Being associates is an equivalence relation: it is

- reflexive since $a = a1$ and 1 is a unit;
- symmetric since, if $b = au$, then $a = bu^{-1}$, and u^{-1} is a unit;
- transitive since, if $b = au$ and $c = bv$ where u and v are units, then $c = a(uv)$, and uv is a unit.

Here we have invoked the three parts of the lemma above about units.

For example, in the ring $\mathbb{Z}/12\mathbb{Z}$, the associate classes are

$$\{0\}, \quad \{1,5,7,11\}, \quad \{2,10\}, \quad \{3,9\} \quad \{4,8\} \quad \{6\}.$$

For example, the associate class containing 2 consists of 2, $2 \cdot 5 = 10$, $2 \cdot 7 = 2$, and $2 \cdot 11 = 10$.

Now we can define greatest common divisors properly.

Let R be an integral domain. (Remember: this means that R is a commutative ring with identity and has no divisors of zero.) We say that *a divides b* in R (written as usual as $a \mid b$) if there exists $x \in R$ with $b = ax$. Notice that every element divides 0, whereas 0 doesn't divide anything else except 0. Also, 1 divides any element of R, but the only elements which divide 1 are the units of R. [Check all these claims!]

Proposition 2.21 *In an integral domain R, two elements a and b are associates if and only if $a \mid b$ and $b \mid a$.*

Proof Suppose that a and b are associates. Then $b = au$ for some unit u, so $a \mid b$. Also $a = bu^{-1}$, so $b \mid a$.

Conversely, suppose that $a \mid b$ and $b \mid a$. If $a = 0$, then also $b = 0$ and a, b are associates. So suppose that $a \neq 0$. Then there are elements x and y such that $b = ax$ and $a = by$. We have $axy = a$, so $a(1 - xy) = 0$. Since R is an integral domain and $a \neq 0$, we must have $1 - xy = 0$, or $xy = 1$. So x and y are units, and a and b are associates.

Now we say that d is a *greatest common divisor* of a and b if

- $d \mid a$ and $d \mid b$;
- if e is any element such that $e \mid a$ and $e \mid b$, then $e \mid b$.

We abbreviate "greatest common divisor" to gcd.

Notice that, in general, "greatest" does not mean "largest" in any obvious way. Both 6 and -6 are greatest common divisors of 12 and 18 in $\mathbb{Z}$, for example.

Proposition 2.22 *If d is a gcd of two elements a, b in an integral domain R, then another element d' is a gcd of a and b if and only if it is an associate of d.*

Proof Suppose first that d and d' are both gcds of a and b. Then $d' \mid d$ and $d \mid d'$ (using the second part of the definition), so that d and d' are associates.

Conversely, suppose that d is a gcd of a and b (say $a = dx$ and $b = dy$), and d' an associate of d, say $d' = du$ for some unit u. Then

- $a = d'u^{-1}x$ and $b = d'u^{-1}y$, so $d' \mid a$ and $d' \mid b$;
- suppose that $e \mid a$ and $e \mid b$. Then $e \mid d$, say $d = ez$; so we have $d' = eu^{-1}z$ and $e \mid d'$.

Thus d' is a gcd of a and b.

Thus we can say: the greatest common divisor of a and b, *if it exists*, is "unique up to associate", that is, any two gcds are associates. We use the notation $\gcd(a, b)$ to denote some (unspecified) greatest common divisor. In the integers, we can make the convention that we choose the non-negative element of the associate pair as the gcd.

2.4.2 Unique factorisation domains

We are interested in the property of "unique factorisation" of integers, that is, any integer other than $0, +1, -1$ can be uniquely factorised into primes. Of course, the factorisation is not quite unique, for two reasons:

(a) the multiplication is commutative, so we can change the order: $6 = 2 \cdot 3 = 3 \cdot 2$.

(b) we will see that -2 and -3 also count as "primes", and $6 = 2 \cdot 3 = (-2) \cdot (-3)$.

By convention, 1 is not a prime, since it divides everything. The same holds for -1 (and only these two integers, since they are the only units in $\mathbb{Z}$.) Accordingly, we will specify that irreducible elements (the analogue of primes in a general domain) should not be zero or units, and that we only try to factorise elements which are not zero or a unit.

So we make the following definitions. Let R be an integral domain.

(a) An element $p \in R$ is *irreducible* if p is not zero or a unit, but whenever $p = ab$, then one of a and b is a unit (and the other therefore an associate of p).

(b) R is a *unique factorisation domain* if it has the following properties:

- every element $a \in R$ which is not zero or a unit can be written as a product of irreducibles;
- if $p_1, \ldots, p_m, q_1, \ldots, q_n$ are irreducibles and
$$p_1 p_2 \cdots p_m = q_1 q_2 \cdots q_n,$$
then $m = n$ and, after possibly permuting the factors in one product, p_i and q_i are associates for $i = 1, \ldots, m$.

Note that, if an element p is irreducible, then so is every associate of p. If the second condition in the definition of a unique factorisation holds, we say that "factorisation is unique up to order and associates". As we saw, this is the best we can expect in terms of unique factorisation!

The ring $\mathbb{Z}$ is a unique factorisation domain; so is the ring $F[x]$ of polynomials over any field F. We will prove these things later on; we will see that it is the Euclidean algorithm which is crucial to the proof, and the integers and polynomials over a field both have a Euclidean algorithm.

Note that, to decide whether a ring is a unique factorisation domain, we have first to check that it really is an integral domain, and second to find all the units (so that we know when two elements are associates).

Example Here is an example of a ring which is an integral domain but not a unique factorisation domain. Let

$$R = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}.$$

We show first that R is a subring of $\mathbb{C}$. Take two elements of R, say $r = a + b\sqrt{-5}$ and $s = c + d\sqrt{-5}$, with $a, b, c, d \in \mathbb{Z}$. Then

$$\begin{aligned} r - s &= (a - c) + (b - d)\sqrt{-5} \in R, \\ rs &= (ac - 5bd) + (ad + bc)\sqrt{-5} \in R, \end{aligned}$$

since $a - c, b - d, ac - 5bc, ad + bc \in \mathbb{Z}$. So the Subring Test applies.

R is clearly an integral domain: there do not exist two nonzero complex numbers whose product is zero.

What are the units of R? To answer this, we use the fact that $|a + b\sqrt{-5}|^2 = a^2 + 5b^2$. Now suppose that $a + b\sqrt{-5}$ is a unit, say

$$(a + b\sqrt{-5})(c + d\sqrt{-5}) = 1.$$

Taking the modulus and squaring gives

$$(a^2 + 5b^2)(c^2 + 5d^2) = 1.$$

So $a^2 + 5b^2 = 1$ (it can't be -1 since it is positive). The only solution is $a = \pm 1$, $b = 0$. So the only units are ± 1, and so r is associate only to r and $-r$.

Now we show that 2 is irreducible. Suppose that

$$2 = (a + b\sqrt{-5})(c + d\sqrt{-5}).$$

Taking the modulus squared again gives

$$4 = (a^2 + 5b^2)(c^2 + 5d^2).$$

So $a^2 + 5b^2 = 1$, 2 or 4. But the equation $a^2 + 5b^2 = 2$ has no solution, while $a^2 + 5b^2 = 1$ implies $a = \pm 1$, $b = 0$, and $a^2 + 5b^2 = 4$ implies $c^2 + 5d^2 = 1$, so that $c = \pm 1$, $d = 0$. So the only factorisations are

$$2 = 2 \cdot 1 = 1 \cdot 2 = (-2) \cdot (-1) = (-2) \cdot (-1) :$$

in each case, one factor is a unit and the other is an associate of 2.

In a similar way we can show that 3, $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$ are irreducible.

Now consider the factorisations

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

These are two factorisations into irreducibles, which are not equivalent up to order and associates. So R is not a unique factorisation domain!

2.4.3 Principal ideal domains

Let R be a commutative ring with identity. We denote by aR, or by $\langle a \rangle$, the set $\{ar : r \in R\}$ of all elements divisible by a.

Lemma 2.23 *$\langle a \rangle$ is an ideal of R containing a, and if I is any ideal of R containing a then $\langle a \rangle \subseteq I$.*

Proof We apply the Ideal Test. If $ar_1, ar_2 \in \langle a \rangle$, then

$$ar_1 - ar_2 = a(r_1 - r_2) \in \langle a \rangle.$$

Also, if $ar \in \langle a \rangle$ and $x \in R$, then

$$(ar)x = a(rx) \in \langle a \rangle.$$

So $\langle I \rangle$ is an ideal.

Since R has an identity element 1, we have $a = a1 \in \langle a \rangle$.

Finally, if I is any ideal containing a, then (by definition of an ideal) we have $ar \in I$ for any $r \in R$; that is, $\langle a \rangle \subseteq I$.

Lemma 2.24 *Let R be an integral domain. Then $\langle a \rangle = \langle b \rangle$ if and only if a and b are associates.*

Proof $\langle a \rangle = \langle b \rangle$ means, by definition, that each of a and b is a multiple of the other, that is, they are associates.

We call $\langle a \rangle$ the *ideal generated by a* and say that it is a *principal ideal*.

More generally, if $a_1, \ldots, a_n \in R$ (where R is a commutative ring with identity, then we let

$$\langle a_1, \ldots, a_n \rangle = \{r_1 a_1 + \cdots + r_n a_n : r_1, \ldots, r_n \in R\}.$$

Then it can be shown, just as above, that $\langle a_1, \ldots, a_n \rangle$ is an ideal of R containing $a_1, \ldots, a_n$, and that any ideal which contains these elements must contain $\langle a_1, \ldots, a_n \rangle$. We call this the *ideal generated by $a_1, \ldots, a_n$*.

A ring R is a *principal ideal domain* if *every* ideal is principal. We will see later that $\mathbb{Z}$ is a principal ideal domain.

Proposition 2.25 *Let R be a principal ideal domain. Then any two elements of R have a greatest common divisor; in fact, $d = \gcd(a, b)$ if and only if $\langle a, b \rangle = \langle d \rangle$.*

Proof Suppose that R is a principal ideal domain. Then $\langle a, b \rangle$, the ideal generated by a and b, is a principal ideal, so it is equal to $\langle d \rangle$, for some $d \in R$. Now we claim that $d = \gcd(a, b)$.

- $a \in \langle d \rangle$, so $d \mid a$. Similarly $d \mid b$.
- Also, $d \in \langle a, b \rangle$, so $d = ua + vb$ for some $u, v \in R$. Now suppose that $e \mid a$ and $e \mid b$, say $a = ep$ and $b = eq$. Then $d = ua + vb = e(up + vq)$, so that $e \mid d$.

The claim is proved.

Since any two gcds of a and b are associates, and any two generators of $\langle a, b \rangle$ are associates, the result is proved.

Example The ring $\mathbb{Z}$ is a principal ideal domain. That means that the only ideals in $\mathbb{Z}$ are the sets $\langle n \rangle = n\mathbb{Z}$, for $n \in \mathbb{Z}$. We will deduce this from a more general result in the next section.

Now it is the case that any principal ideal domain is a unique factorisation domain. We will not prove all of this. The complete proof involves showing two things: any element which is not zero or a unit can be factorised into irreducibles; and any two factorisations of the same element differ only by order and associates. We will prove the second of these two assertions. See the appendix to this chapter for comments on the first.

Lemma 2.26 *Let R be a principal ideal domain; let p be irreducible in R, and $a, b \in R$. If $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof Suppose that $p \mid ab$ but that p does not divide a. Then we have $\gcd(a, p) = 1$, and so there exist $u, v \in R$ with $1 = ua + vp$. So $b = uab + vpb$. But $p \mid uab$ by assumption, and obviously $p \mid vpb$; so $p \mid b$, as required.

This lemma clearly extends. If p is irreducible and divides a product $a_1 a_2 \cdots a_n$, then p must divide one of the factors. For either $p \mid a_1$ or $p \mid a_2 \cdots a_n$; in the latter case, proceed by induction.

Theorem 2.27 *Let R be a principal ideal domain, and suppose that*

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

where $p_1, \ldots, p_m, q_1, \ldots, q_n$ are irreducible. Then $m = n$ and, after possibly permuting the factors, p_i and q_i are associates for $i = 1, \ldots, m$.

Proof Obviously p_1 divides $q_1 \cdots q_n$, so p_1 must divide one of the factors, say $p_1 \mid q_i$. Since p_1 and q_i are irreducible, they must be associates. By permuting the order of the qs and adjusting them by unit factors, we can assume that $p_1 = q_1$. Then $p_2 \cdots p_m = q_2 \cdots q_n$, and we proceed by induction.

Example Here is an example of an integral domain which is not a principal ideal domain. Consider the ring $R = \mathbb{Z}[x]$ of polynomials over the integers. Let I be the set of all such polynomials whose constant term is even. Then I is an ideal in R: if f and g are polynomials with even constant term, then so is $f - g$, and so is fh for any polynomial h. But I is not a principal ideal. For I contains both the constant polynomial 2 and the polynomial x of degree 1. If $I = \langle a \rangle$, then a must divide both 2 and x, so $a = \pm 1$. But $\pm 1 \notin I$.

The polynomials 2 and x are both irreducible in R, and so their gcd is 1. But 1 cannot be written in the form $2u + xv$ for any polynomials u and v.

The ring $\mathbb{Z}[x]$ is a unique factorisation domain (see the Appendix to this chapter).

2.4.4 Euclidean domains

Any two integers have a greatest common divisor, and we can use the Euclidean algorithm to find it. You may also have seen that the Euclidean algorithm works for polynomials. We now give the algorithm in a very general form.

Let R be an integral domain. A *Euclidean function* on R is a function d from the set $R \setminus \{0\}$ (the set of non-zero elements of R) to the set $\mathbb{N}$ of non-negative integers satisfying the two conditions

(a) for any $a, b \in R$ with $a, b \neq 0$, we have $d(ab) \geq d(a)$;

(b) for any $a, b \in R$ with $b \neq 0$, there exist $q, r \in R$ such that

- $a = bq + r$;
- either $r = 0$, or $d(r) < d(b)$.

We say that an integral domain is a *Euclidean domain* if it has a Euclidean function.

Example Let $R = \mathbb{Z}$, and let $d(a) = |a|$ for any integer a.

Example Let $R = F[x]$, the ring of polynomials over F, where F is a field. For any non-zero polynomial $f(x)$, let $d(f(x))$ be the degree of the polynomial $f(x)$ (the index of the largest non-zero coefficient).

Both of these examples are Euclidean functions.

(a) In the integers, we have $d(ab) = |ab| = |a| \cdot |b| \geq |a| = d(a)$, since $b \neq 0$. In the polynomial ring $F[x]$, we have

$$d(ab) = \deg(ab) = \deg(a) + \deg(b) \geq \deg(a),$$

since if the leading terms of a and b are $a_n x^n$ and $b_m x^m$ respectively then the leading term of ab is $a_n b_m x^{n+m}$.

(b) In each case this is the "division algorithm": we can divide a by b to obtain a quotient q and remainder r, where r is smaller than the divisor b as measured by the appropriate function d.

You will have seen how to use the Euclidean algorithm to find the greatest common divisor of two integers or two polynomials. The same method works in any Euclidean domain. It goes like this. Suppose that R is a Euclidean domain, with Euclidean function d. Let a and b be any two elements of R. If $b = 0$, then $\gcd(a,b) = a$. Otherwise, proceed as follows. Put $a = a_0$ and $b = a_1$. If a_{i-1} and a_i have been constructed, then

- if $a_i = 0$ then $\gcd(a,b) = a_{i-1}$;
- otherwise, write $a_{i-1} = a_i q + r$, with $r = 0$ or $d(r) < d(a_i)$, and set $a_{i+1} = r$; repeat the procedure for a_i and a_{i+1}.

The algorithm terminates because, as long as $a_i \neq 0$, we have

$$d(a_i) < d(a_{i-1}) < \cdots < d(a_1).$$

Since the values of d are non-negative integers, this chain must stop after a finite number of steps.

To see that the result is correct, note that, if $a = bq + r$, then

$$\gcd(a,b) = \gcd(b,r)$$

(as an easy calculation shows: the common divisors of a and b are the same as the common divisors of b and r. So we have $\gcd(a_{i-1}, a_i) = \gcd(a,b)$ as long as a_i is defined. At the last step, $a_i = 0$ and so $\gcd(a,b) = \gcd(a_{i-1}, 0) = a_{i-1}$.

The algorithm can also be used to express $\gcd(a,b)$ in the form $ua + vb$ for some $u, v \in R$. For a and b themselves are both expressible in this form; and, if $a_{i-1} = u_{i-1}a + v_{i-1}b$ and $a_i = u_i a + v_i b$, then with $a_{i-1} = qa_i + a_{i+1}$, we have

$$a_{i+1} = a_{i-1} - qa_i = (u_{i-1} - qu_i)a + (v_{i-1} - qv_i)b.$$

Example Find $\gcd(204, 135)$. We have

$$\begin{aligned} 204 &= 135 \cdot 1 + 69, \\ 135 &= 69 \cdot 1 + 66, \\ 69 &= 66 \cdot 1 + 3, \\ 66 &= 3 \cdot 22, \end{aligned}$$

so $\gcd(204, 135) = 3$. To express $3 = 204u + 135v$, we have

$$\begin{aligned} 69 &= 204 \cdot 1 - 135 \cdot 1, \\ 66 &= 135 - 69 = 135 \cdot 2 - 204 \cdot 1, \\ 3 &= 69 - 66 = 204 \cdot 2 - 135 \cdot 3. \end{aligned}$$

We will show that a Euclidean domain is a unique factorisation domain. First we need one lemma. Note that, if a and b are associates, then $b = au$, so $d(b) \geq d(a)$, and also $a = bu^{-1}$, so $d(a) \geq d(b)$; so we have $d(a) = d(b)$.

Lemma 2.28 *Let R be a Euclidean domain. Suppose that a and b are non-zero elements of R such that $a \mid b$ and $d(a) = d(b)$. Then a and b are associates.*

Proof Let $a = bq + r$ for some q, r, as in the second part of the definition. Suppose that $r \neq 0$. Now $b = ac$ for some element c; so $a = acq + r$. Thus, $r = a(1 - cq)$, and since $r \neq 0$ we have $d(r) \geq d(a)$, contrary to assumption. So $r = 0$. Then $b \mid a$; since we are given that $a \mid b$, it follows that a and b are associates.

Theorem 2.29 *(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, and let I be an ideal in R. If $I = \{0\}$, then certainly $I = \langle 0 \rangle$ and I is principal. So suppose that I is not $\{0\}$. Since the values of $d(x)$ for $x \in I$ are non-negative integers, there must be a smallest value, say $d(a)$. We will claim that $I = \langle a \rangle$.

First, take $b \in \langle a \rangle$, say $b = ax$. Then $b \in I$, by definition of an ideal.

Next, take $b \in I$. Use the second part of the definition of a Euclidean function to find elements q and r such that $b = aq + r$, with either $r = 0$ or $d(r) < d(a)$. Suppose that $r \neq 0$. Then $b \in I$ and $aq \in I$, so $r = b = aq \in I$; but $d(r) < d(a)$ contradicts the fact that $d(a)$ was the smallest value of the function d on the non-zero elements of I. So the supposition is impossible; that is, $r = 0$, and $b = aq \in \langle a \rangle$.

So $I = \langle a \rangle$ is a principal ideal.

(b) Again let R be a Euclidean domain. We show that any nonzero non-unit of R can be factorised into irreducibles. We showed in the last section that the factorisation is unique (because R is a principal ideal domain)

Choose any element $a \in R$ such that $a \neq 0$ and a is not a unit. We have to show that a can be factorised into irreducibles. The proof is by induction on $d(a)$; so we can assume that any element b with $d(b) < d(a)$ has a factorisation into irreducibles.

If a is irreducible, then we have the required factorisation with just one term. So suppose that $a = bc$ where b and c are not units. If $d(b) < d(a)$ and $d(c) < d(a)$ then, by induction, each of b and c has a factorisation into irreducibles; putting these together we get a factorisation of a. So suppose that $d(a) \geq d(b)$. We also have $d(b) \geq d(a)$, by the first property of a Euclidean function; so $d(a) = d(b)$. We also have $b \mid a$; by the Lemma before the Theorem, we conclude that a and b are associates, so that c is a unit, contrary to assumption.

Corollary 2.30 *(a) $\mathbb{Z}$ is a principal ideal domain and a unique factorisation domain.*

(b) For any field F, the ring $F[x]$ of polynomials over F is a principal ideal domain and a unique factorisation domain.

Proof This follows from the theorem since we have seen that these rings are integral domains and have Euclidean functions, and so are Euclidean domains.

2.4.5 Appendix

More is true than we have proved above. You will meet these theorems in the Algebraic Structures II course next term.

The connection between the three types of domain is:

Theorem 2.31

Euclidean domain $\Rightarrow$ principal ideal domain $\Rightarrow$ unique factorisation domain.

We proved most of this: we showed that a Euclidean domain is a principal ideal domain, and that in a principal ideal domain factorisations are unique if they exist. The proof that factorisations into irreducibles always exist in a principal ideal domain is a little harder.

Neither implication reverses. We saw that $\mathbb{Z}[x]$ is not a principal ideal domain, though it is a unique factorisation domain (see below). It is harder to construct a ring which is a principal ideal domain but not a Euclidean domain, though such rings do exist.

Another way to see the increasing strength of the conditions from right to left is to look at greatest common divisors.

- In a unique factorisation domain, any two elements a and b have a greatest common divisor d (which is unique up to associates).
- In a principal ideal domain, any two elements a and b have a greatest common divisor d (which is unique up to associates), and d can be written in the form $d = xa + yb$.

- In a Euclidean domain, any two elements a and b have a greatest common divisor d (which is unique up to associates), and d can be written in the form $d = xa + yb$; moreover, the gcd, and the elements x and y, can be found by the Euclidean algorithm.

You will also meet the theorem known as *Gauss's Lemma*:

Theorem 2.32 *If R is a unique factorisation domain, then so is $R[x]$.*

This result shows that $\mathbb{Z}[x]$ is a unique factorisation domain, as we claimed above.

2.5 Fields

As you know from linear algebra, fields form a particularly important class of rings, since in linear algebra the scalars are always taken to form a field.

Although the ring with a single element 0 would technically qualify as a field according to our definition, we always rule out this case. Thus,

A field must have more than one element.

Another way of saying the same thing is that, in a field, we must have $1 \neq 0$. (If there is any element $x \neq 0$ in a ring with identity, then $1 \cdot x = x \neq 0 = 0 \cdot x$, and so $1 \neq 0$.)

The "standard" examples of fields are the rational, real and complex numbers, and the integers mod p for a prime number p.

In this chapter, we will see how new fields can be constructed. The most important method of construction is *adjoining a root of a polynomial*. The standard example of this is the construction of $\mathbb{C}$ by adjoining the square root of -1 (a root of the polynomial $x^2 + 1 = 0$) to $\mathbb{R}$. We will also see that finite fields can be constructed in this way.

Also we can build fields as *fields of fractions*; the standard example is the construction of the rationals from the integers.

2.5.1 Maximal ideals

In this chapter, R always denotes a commutative ring with identity. As above, we assume that the identity element 1 is different from the zero element 0: that is, $0 \neq 1$.

An ideal I of R is said to be *proper* if $I \neq R$. An ideal I is *maximal* if $I \neq R$ and there does not exist an ideal J with $I \subset J \subset R$; that is, any ideal J with $I \subseteq J \subseteq R$ must satisfy $J = I$ or $J = R$.

Lemma 2.33 *Let R be a commutative ring with identity. Then R is a field if and only if it has no ideals except $\{0\}$ and R.*

Proof If $u \in R$ is a unit, then the only ideal containing u is the whole ring R. (For, given any ideal I with $u \in I$, and any $r \in R$, we have $r = u(u^{-1}r) \in I$, so $I = R$.) If R is a field, then every non-zero element is a unit, and so any ideal other than $\{0\}$ is R.

Conversely, suppose that the only ideals are 0 and R. We have to prove that multiplicative inverses exist (axiom (M3)). Take any element $a \in R$ with $a \neq 0$. Then $\langle a \rangle = R$, so $1 \in \langle a \rangle$. This means that there exists $b \in R$ with $ab = 1$, so $b = a^{-1}$ as required.

Proposition 2.34 *Let F be a commutative ring with identity, and I a proper ideal of R. Then R/I is a field if and only if I is a maximal ideal.*

Proof By the Second Isomorphism Theorem, ideals of R/I correspond to ideals of R containing I. Thus, I is a maximal ideal if and only if the only ideals of R/I are zero and the whole ring, that is, R/I is a field (by the Lemma).

Proposition 2.35 *Let R be a principal ideal domain, and $I = \langle a \rangle$ an ideal of R. Then*

(a) $I = R$ if and only if a is a unit;

(b) I is a maximal ideal if and only if a is irreducible.

Proof (a) If a is a unit, then for any $r \in R$ we have $r = a(a^{-1}r) \in \langle a \rangle$, so $\langle a \rangle = R$. Conversely, if $\langle a \rangle = R$, then $1 = ab$ for some $b \in R$, and a is a unit.

(b) Since R is a PID, any ideal containing $\langle a \rangle$ has the form $\langle b \rangle$ for some $b \in R$. Moreover, $\langle a \rangle \subseteq \langle b \rangle$ if and only if $b \mid a$. So $\langle a \rangle$ is maximal if and only if, whenever $b \mid a$, we have either b is a unit (so $\langle b \rangle = R$) or b is an associate of a (so $\langle b \rangle = \langle a \rangle$.

Corollary 2.36 *$\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime.*

Proof $\mathbb{Z}$ is a principal ideal domain, and irreducibles are just the prime integers.

The field $\mathbb{Z}/p\mathbb{Z}$, for a prime number p, is often denoted by $\mathbb{F}_p$.

2.5.2 Adding the root of a polynomial

The other important class of principal ideal domains consists of the polynomial rings over fields. For these, Propositions 2.34 and 2.35 give the first part of the following result.

Proposition 2.37 *Let F be a field and $f(x)$ an irreducible polynomial over F. Then $K = F[x]/\langle f(x) \rangle$ is a field. Moreover, there is an isomorphism from F to a subfield of K; and, if α denotes the coset $\langle f(x) \rangle + x$, then we have the following, where n is the degree of $f(x)$, and we identify an element of F with its image under the isomorphism:*

(a) every element of k can be uniquely written in the form

$$c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_{n-1}\alpha^{n-1};$$

(b) $f(\alpha) = 0$.

Before proving this, we notice that this gives us a construction of the complex numbers; Let $F = \mathbb{R}$, and let $f(x) = x^2 + 1$ (this polynomial is irreducible over $\mathbb{R}$). Use the notation i instead of α for the coset $\langle f(x) \rangle + x$. Then we have $n = 2$, and the two parts of the proposition tell us that

(a) every element of K can be written uniquely as $a + b\mathrm{i}$, where $a, b \in \mathbb{R}$;

(b) $\mathrm{i}^2 = -1$.

Thus, $K = \mathbb{R}[x]/\langle x^2 + 1 \rangle$ is the field $\mathbb{C}$. The general theory tells us that this construction of $\mathbb{C}$ does produce a field; it is not necessary to check all the axioms.

Proof (a) Let I denote the ideal $\langle f(x) \rangle$. Remember that the elements of the quotient ring $F[x]/I$ are the cosets of I in $F[x]$. The isomorphism θ from F to $K = F[x]/I$ is given by

$$a\theta = I + a \quad \text{for} \quad a \in F.$$

Clearly θ is one-to-one; for if $a\theta = b\theta$, then $b - a \in I$, but I consists of all multiples of the irreducible polynomial $f(x)$, and cannot contain any constant polynomial except 0, so $a = b$. It is routine to check that θ preserves addition and multiplication. From now on, we identify a with the coset $I + a$, and regard F as a subfield of $F[x]/I$.

Let $g(x) \in F[x]$. Then by the Euclidean algorithm we can write

$$g(x) = f(x)q(x) + r(x),$$

where $r(x) = 0$ or $r(x)$ has degree less than n. Also, since $g(x) - r(x)$ is a multiple of $f(x)$, it belongs to I, and so the cosets $I + g(x)$ and $I + r(x)$ are equal. In other words, every coset of I in $F[x]$ has a coset representative with degree less than n (possibly zero). This coset representative is unique, since the difference between any two coset representatives is a multiple of $f(x)$.

Now let $r(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{n-1}x^{n-1}$. We have

$$\begin{aligned} I + r(x) &= I + (c_0 + c_1x + c_2x^2 + \cdots + c_{n-1}x^{n-1}) \\ &= (I + c_0) + (I + c_1)(I + x) + (I + c_2)(I + x)^2 + \cdots + (I + c_{n-1})(I + x)^{n-1}) \\ &= c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_{n-1}\alpha^{n-1}. \end{aligned}$$

Here, in the second line, we use the definition of addition and multiplication of cosets, and in the third line we put $I + x = \alpha$ and use our identification of $I + c = c\theta$ with c for $c \in F$.

So we have the required representation. Clearly it is unique.

(b) As before, if $f(x) = a_0 + a_1x + \cdots + a_nx^n$, we have $I + f(x) = I$ (since $f(x) \in I$), and so

$$\begin{aligned} 0 &= I + 0 \\ &= I + (a_0 + a_1x + \cdots + a_nx^n) \\ &= (I + a_0) + (I + a_1)(I + x) + \cdots + (I + a_n)(I + x)^n \\ &= a_0 + a_1\alpha + \cdots + a_n\alpha^n \\ &= f(\alpha). \end{aligned}$$

2.5.3 Finite fields

Suppose that $f(x)$ is an irreducible polynomial of degree n over the field $\mathbb{F}_p$ of integers mod p. Then $K = \mathbb{F}_p[x]/\langle f(x)\rangle$ is a field, by Proposition 2.37. According to that proposition, its elements can be written uniquely in the form

$$c_0 + c_1\alpha + \cdots + c_{n-1}\alpha^{n-1}$$

for $c_0, \ldots, c_{n-1} \in \mathbb{F}_p$. There are p choices for each of the n coefficients $c_0, c_1, \ldots, c_{n-1}$, giving a total of p^n elements altogether. Thus:

Proposition 2.38 *Let $f(x)$ be an irreducible polynomial of degree n over $\mathbb{F}_p$. Then $K = \mathbb{F}_p[x]/\langle f(x)\rangle$ is a field containing p^n elements.*

Example Let $p = 2$ and $n = 2$. The coefficients of a polynomial over $\mathbb{F}_2$ must be 0 or 1, and so there are just four polynomials of degree 2, namely x^2, $x^2 + 1$, $x^2 + x$ and $x^2 + x + 1$. We have

$$x^2 = x \cdot x, \qquad x^2 + x = x \cdot (x + 1), \qquad x^2 + 1 = (x + 1) \cdot (x + 1)$$

(remember that $1 + 1 = 0$ in $\mathbb{F}_2$!), and so the only irreducible polynomial is $x^2 + x + 1$. Thus, there is a field consisting of the four elements $0, 1, \alpha, 1 + \alpha$, in which $\alpha^2 + \alpha + 1 = 0$, that is, $\alpha^2 = 1 + \alpha$ (since $-1 = +1$ in $\mathbb{F}_2$!) The addition and multiplication tables are easily found (with $\beta = 1 + \alpha$) to be

$+$	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

We have, for example,

$$\begin{aligned} \alpha + \beta &= \alpha + 1 + \alpha = 1, \\ \alpha\beta &= \alpha(1 + \alpha) = \alpha + \alpha^2 = 1, \\ \beta^2 &= (1 + \alpha)^2 = 1 + \alpha = \beta. \end{aligned}$$

The basic facts about finite fields were one of the discoveries of Évariste Galois, the French mathematician who was killed in a duel in 1832 at the age of 19. Most of his mathematical work, which is fundamental for modern algebra, was not published until fifteen years after his death, but the result on finite fields was one of the few papers published during his lifetime.

Galois proved the following theorem:

Theorem 2.39 *The number of elements in a finite field is a power of a prime. For any prime power p^n, there is a field with p^n elements, and any two finite fields with the same number of elements are isomorphic.*

We commemorate Galois by using the term *Galois field* for finite field. If $q = p^n$, then we often denote the field with q elements by $\mathrm{GF}(q)$. Thus the field on the preceding page is $\mathrm{GF}(4)$. (Note that $\mathrm{GF}(4)$ is *not* the same as $\mathbb{Z}/4\mathbb{Z}$, the integers mod 4, which is not a field!)

2.5.4 Field of fractions

In this section we generalise the construction of the rational numbers from the integers. [This section and the two following were not covered in the lectures, but you are encouraged to read them for interest.]

Theorem 2.40 *Let R be an integral domain. Then there is a field F such that*

(a) R is a subring of F;

(b) every element of F has the form ab^{-1}, for $a,b \in R$ and $b \neq 0$.

The field F is called the *field of fractions* of R, since every element of F can be expressed as a fraction a/b.

We will build F as the set of all fractions of this form. But we have to answer two questions?

- When are two fractions equal?
- How do we add and multiply fractions?

Thus, we start with the set X consisting of all ordered pairs (a,b), with $a,b \in R$ and $b \neq 0$. (That is, $X = R \times (R \setminus \{0\})$.) The ordered pair (a,b) will "represent" the fraction a/b. So at this point we have to answer the first question above: when does $a/b = c/d$? Multiplying up by bd, we see that this holds if and only if $ad = bc$. Thus, we define a relation $\sim$ on X by the rule

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

We have to show that this is an equivalence relation.

reflexive: $ab = ba$, so $(a,b) \sim (a,b)$.

symmetric: If $(a,b) \sim (c,d)$, then $ad = bc$, so $cb = da$, whence $(c,d) \sim (a,b)$.

transitive: Suppose that $(a,b) \sim (c,d)$ and $(c,d) \sim (e,f)$. Then $ad = bc$ and $cf = de$. So $adf = bcf = bde$. This means that $d(af - be) = 0$. But $d \neq 0$ and R is an integral domain, so we conclude that $af = be$, so that $(a,b) \sim (e,f)$.

Now we let F be the set of equivalence classes of the relation $\sim$. We write the equivalence class containing (a,b) as a/b. Thus we do indeed have that $a/b = c/d$ if and only if $ad = bc$.

Now we define addition and multiplication by the "usual rules":

- $(a/b) + (c/d) = (ad + bc)/(bd)$;
- $(a/b)(c/d) = (ac)/(bd)$.

(To see where these rules come from, just calculate these fractions in the usual way!) Again, since $b \neq 0$ and $d \neq 0$, we have $bd \neq 0$, so these operations make sense. We still have to show that they are well-defined, that is, a different choice of representatives would give the same result. For addition, this means that, if $(a,b) \sim (a',b')$ and $(c,d) \sim (c',d')$, then $(ad+bc,bd) \sim (a'd'+b'c',b'd')$. Translating, we have to show that

$$\text{if } ab' = ba' \text{ and } cd' = dc', \text{ then } (ad+bc)b'd' = bd(a'd'+b'c'),$$

a simple exercise. The proof for multiplication is similar.

Now we have some further work to do. We have to show that

- F, with addition and multiplication defined as above, is a field;
- the map θ defined by $a\theta = a/1$ is a homomorphism from R to F, with kernel $\{0\}$ (so that R is isomorphic to the subring $\{a/1 : a \in R\}$ of F).

These are fairly straightforward to prove, and their proof finishes the theorem.

2.5.5 Appendix: Simple rings

We saw at the start of this chapter (Lemma 2.33) that, if R is a commutative ring with identity having no ideals except the trivial ones, then R is a field. You might think that, if we simply leave out the word "commutative", then we obtain a characterisation of division rings. Unfortunately this is not so. The material here is not part of the course; you can find a proof in the course textbook if you are interested. Let R be a ring with identity. We say that R is a *simple ring* if the only ideals in R are $\{0\}$ and R. Then every division ring (and in particular every field) is a simple ring, and our earlier argument shows that a commutative simple ring is a field. But we have the following fact:

Theorem 2.41 *Let R be a simple ring (with identity). Then the ring $M_n(R)$ of $n \times n$ matrices over R is a simple ring.*

In particular, the ring of $n \times n$ matrices over a field F is a simple ring, although it is not commutative and is not a division ring for $n > 1$.