(R1)	for all x and y in R, $x + y$ is in R	closure under addition;
(R2)	for all x, y and z in R, $x + (y + z) = (x + y) + z$	associativity of addition;
(R3)	there is an element, 0, in R such that for all x in R $x + 0 = x = 0 + x$	existence of zero element;
(R4)	for every element x of R there is an element $-x$ in R such that $x + (-x) = 0 = (-x) + x$	existence of negatives;
(R5)	for all x and y in R, $x + y = y + x$	commutativity of addition;
(R6)	for all x, y in R, xy is in R	closure under multiplication;
(R7)	for all x, y and z in R, $x(yz) = (xy)z$	associativity of multiplication;
(R8)	for all x, y and z in R, $x(y + z) = xy + xz$, and $(x + y)z = xz + yz$	distributivity.

The first five axioms say that R is an Abelian group under addition: axioms (R6) and (R7) say that R is a semigroup under multiplication. The eighth axiom is the one that says how the two operations are linked. The above list of axioms can therefore be summarised by saying that a ring is a set, equipped with operations called addition and multiplication, which is an additive Abelian group, is also a multiplicative semigroup, and in which multiplication distributes over addition.

Example 1 The set $\mathbb{Z}$ of integers with the usual addition and multiplication is a ring. Notice that this ring has an identity element, 1, with respect to multiplication, and also has commutative multiplication.

The set $2\mathbb{Z}$ of all even integers also is a ring, but it has no multiplicative identity: clearly 0 is not a multiplicative identity and if $n = 2m$ ($m \in \mathbb{Z}$) were an identity in this ring it would, in particular, be idempotent and so we would have $2m = (2m)^2 = 4m^2$ and hence $2m = 1$, contrary to m being an integer.

Example 2 The set $M_2(\mathbb{R})$ of 2×2 matrices with real coefficients is a ring under matrix addition and multiplication. This ring also has a multiplicative identity (the 2×2 identity matrix), but the multiplication is not commutative.

Many of the common examples of rings exhibit various significant special properties. Recall from Section 1.4 that an element x is a **zero-divisor** if x is not zero and if there is a non-zero element y with $xy = 0$. There are zero-divisors in Example 2 above: e.g.

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Example 3 The set $\mathbb{Z}_n$ of congruence classes modulo n is a ring under the usual addition and multiplication. As we saw in Section 1.4 (Theorem 1.4.3 and Corollary 1.4.5), this set has zero-divisors unless n is a prime, in which case every non-zero element has a multiplicative inverse.

Example 4 Define $\mathbb{Z}[\sqrt{2}]$ to be the set of all real numbers of the form $a + b\sqrt{2}$ where a and b are integers. Then, equipped with the operations of addition and multiplication which are inherited from $\mathbb{R}$, this is a ring. Specifically, the operations are

$$(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2},$$
$$(a + b\sqrt{2}) \times (c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2}.$$

We have just noted that the set is closed under the operations; clearly the set is closed under taking additive inverses (i.e. negatives); the other properties – associativity, distributivity, etc. – are inherited from $\mathbb{R}$ (they are true in $\mathbb{R}$ so certainly hold in the smaller subset $\mathbb{Z}[\sqrt{2}]$).

A similar example is $\mathbb{Z}[\mathrm{i}]$ (where $\mathrm{i}^2 = -1$): we obtain a subset of the ring $\mathbb{C}$ of complex numbers which is a ring in its own right.

We will consider rings of polynomials in Chapter 6. Also see Example 1 on p. 191.

Definition A **field** is a set F equipped with two operations ('addition' and 'multiplication'), under which it is a commutative ring with identity element $1 \neq 0$ in which every non-zero element has a multiplicative inverse.

Example 1 For any prime p, the set $\mathbb{Z}_p$ is a field by Corollary 1.4.6.

Example 2 The sets $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ all are fields. The study of more general fields arose from the work of Galois (see below and the historical notes at the end of this section) and from that of Dedekind and Kronecker on number theory. The abstract study of fields was initiated by Weber (c. 1893) and Hensel and Steinitz made fundamental contributions at the beginning of the twentieth century.

Fields arose in Galois' work as follows. Given a polynomial $p(x)$ with rational coefficients, in the indeterminate x, the 'Fundamental Theorem of Algebra' says that it can be factorised completely into linear factors with complex coefficients. If we take the roots of this polynomial, we can adjoin them to the field $\mathbb{Q}$ of rational numbers and form the smallest extension field of $\mathbb{Q}$ containing them. The groups that Galois introduced are intimately connected with such

extension fields of the rationals. (The connection is studied under the name 'Galois Theory'.)

For an example of this adjunction of roots, consider Example 3 below.

Example 3 The ring $\mathbb{Z}[\sqrt{2}]$ defined above is not a field but, if we define the somewhat larger set $\mathbb{Q}[\sqrt{2}]$ to be the set of all real numbers of the form $a + b\sqrt{2}$ where a and b are rational numbers, then we do obtain a field. The main point to be checked is that this set does contain a multiplicative inverse for each of its non-zero elements (checking the other axioms for a field is left as an exercise). So let $a + b\sqrt{2}$ be non-zero (thus at least one of a, b is non-zero and hence $a^2 - 2b^2 \neq 0$ since $\sqrt{2}$ is not rational). Then $(a + b\sqrt{2}) \times (c + d\sqrt{2}) = 1$ where $c = a/(a^2 - 2b^2)$ and $d = -b/(a^2 - 2b^2)$.

Observe, in connection with the comments in Example 2 above, that the polynomial $x^2 - 2$ factorises if we allow coefficients from the field $\mathbb{Q}[\sqrt{2}]$: $x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2})$. But it does not factorise over $\mathbb{Q}$ since $\sqrt{2}$ is not a rational number. We may think of $\mathbb{Q}[\sqrt{2}]$ as having been obtained from $\mathbb{Q}$ by adjoining the roots $\sqrt{2}$ and $-\sqrt{2}$ then closing under addition, multiplication (and forming inverses of non-zero elements) so as to obtain the smallest field containing $\mathbb{Q}$ together with the roots of $x^2 - 2$.

As with groups, the axioms for our various algebraic systems have significant consequences. For example, the zero element in any ring is unique. It follows also (cf. proof of Corollary 1.4.5) that a field has no zero-divisors. We give an example of the way in which some of these consequences may be deduced.

Theorem 4.4.1 *Let R be any ring and let x be an element of R. Then*:

$$x0 = 0 = 0x.$$

Proof Write y for $x0$. Then

$$\begin{aligned} y + y &= x0 + x0 = x(0 + 0) && \text{using (R8)} \\ &= x0 && \text{using (R3)} \\ &= y. \end{aligned}$$

Thus $y + y = y$. Add $-y$ (which exists by condition (R4)) to each side of this equation, to obtain (using R2)) $y = 0$, as required. The proof for $0x$ is similar. □

One of the most commonly arising algebraic structures is composed of a field together with an Abelian group on which the field acts in a certain way: the Abelian group is then called a vector space over the field.

Definition Given a field F, a **vector space** V over F is an additive Abelian group which also has a **scalar multiplication**. The scalar multiplication is an operation which takes any $\lambda \in F$ and $v \in V$ and gives an element, written λv, of V. The following axioms are to be satisfied:

(V1) for all v in V and λ in F, λv is an element of V;
(V2) for all v in V and λ, μ in F, $(\lambda\mu)v = \lambda(\mu v)$;
(V3) for all v in V, $1v = v$;
(V4) for all v in V and λ, μ in F, $(\lambda + \mu)v = \lambda v + \mu v$;
(V5) for all u, v in V and λ in F, $\lambda(u + v) = \lambda u + \lambda v$.

The elements of V are called **vectors** and the elements of the field F are called **scalars**. Vector spaces are usually studied in courses or books with titles such as 'Linear Algebra'.

Example The most familiar examples of vector spaces occur when F is the field $\mathbb{R}$ of real numbers. Taking V to be $\mathbb{R}^2$, the set of ordered pairs (x, y) with x and y real numbers, addition and scalar multiplication are defined by

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2), \quad \text{and}$$
$$\lambda(x, y) = (\lambda x, \lambda y).$$

This makes the real plane $\mathbb{R}^2$ into a vector space.

We consider some more well known mathematical objects in the light of the structures we have introduced.

Example 1 Consider the set, $\mathbb{R}[x]$, of polynomials with real coefficients in the variable x. Clearly, we can add two polynomials by adding the coefficients of each power of x, and the result will be in $\mathbb{R}[x]$. For example

$$(x^2 + 3x + \pi) + (-5x^2 + 3) = -4x^2 + 3x + (\pi + 3).$$

The set is an additive Abelian group under this operation. We can also multiply two polynomials by collecting together powers of x. For example

$$(\sqrt{3} \cdot x - 1) \times (175x^2 + x + 1) = 175\sqrt{3} \cdot x^3 + (\sqrt{3} - 175)x^2 + (\sqrt{3} - 1)x - 1.$$

It is straightforward to check that $\mathbb{R}[x]$ is a ring under these operations.

The ring $\mathbb{R}[x]$ is not a field, since the polynomial $x + 1$ (for instance) does not have a multiplicative inverse. However $\mathbb{R}[x]$ is a commutative ring with identity that has no zero-divisors (a ring with these properties is called an **integral domain**).

The set $\mathbb{R}[x]$ has yet more structure: we can define a scalar multiplication by real numbers on the elements of $\mathbb{R}[x]$, by multiplying each coefficient of a given polynomial by a given scalar. For example

$$3\pi \cdot (x^2 + 3x) = 3\pi x^2 + 9\pi x.$$

Then $\mathbb{R}[x]$ becomes a vector space over $\mathbb{R}$. A ring that is at the same time a vector space (over a field K) under the same operation of addition, is known as a (K-)**algebra**. Thus $\mathbb{R}[x]$ is an $\mathbb{R}$-algebra. (Rings of) polynomials will be discussed at length in Chapter 6.

Example 2 Given a prime number p, we consider the set $M_2(\mathbb{Z}_p)$ of 2×2 matrices whose entries are in $\mathbb{Z}_p$. Under the usual addition and multiplication of matrices, this is a ring with identity. However, this ring is not commutative and it does have zero-divisors. Again, we can define a scalar multiplication on $M_2(\mathbb{Z}_p)$ by setting, for $\lambda \in \mathbb{Z}_p$,

$$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix} \text{ (where we are computing modulo } p\text{).}$$

This gives $M_2(\mathbb{Z}_p)$ the structure of a vector space over the field $\mathbb{Z}_p$ and so $M_2(\mathbb{Z}_p)$ is also an algebra (a $\mathbb{Z}_p$ - algebra).

Example 3 Consider the set C of 2×2 matrices of the form $a\mathbf{I} + bY$ where a and b are real numbers and $\mathbf{I}$ and Y are respectively the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Equip this set with the usual matrix addition and multiplication. It is easily checked that this set is an $\mathbb{R}$-algebra.

In fact, we have just given one way of constructing the complex numbers. Regard $a\mathbf{I} + bY$ as being '$a \cdot 1 + b\mathrm{i}$'. You may check that $Y^2 = -\mathbf{I}$ and that these matrices add and multiply in the same way as expressions of the form $a + b\mathrm{i}$ where $\mathrm{i}^2 = -1$. The details of checking all this are left as an exercise for the reader.

Finally in this section, we consider one more type of structure. A **Boolean algebra** is a set B, together with operations '$\wedge$', '$\vee$' and '$\neg$' which will be called 'meet', 'join' and 'complement', such that for all $a, b \in B$ we have $a \wedge b$,

$a \vee b$ and $\neg a$ all in B, together with two distinguished elements, denoted 0 and 1, which are different ($0 \neq 1$). The axioms which must be satisfied are as follows.

Let a, b and c denote elements of B, then

$a \wedge a = a$ and	
$a \vee a = a$	idempotence,
$a \vee \neg a = 1$ and	
$a \wedge \neg a = 0$	complementation,
$a \wedge b = b \wedge a$ and	
$a \vee b = b \vee a$	commutativity,
$a \wedge (b \wedge c) = (a \wedge b) \wedge c$ and	
$a \vee (b \vee c) = (a \vee b) \vee c$	associativity,
$\neg(a \wedge b) = \neg a \vee \neg b$ and	
$\neg(a \vee b) = \neg a \wedge \neg b$	De Morgan laws,
$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$ and	
$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$	distributivity,
$\neg\neg a = a$	double complement,
$a \wedge 1 = a$	property of 1,
$a \vee 0 = a$	property of 0.

We have already seen some examples of Boolean algebras.

Example 1 Let U be any set. Let B be the set of all subsets of U. Equip B with the operations of intersection, union and complement for its Boolean meet, join and complement. Let U play the role of 1 and the empty set that of 0. Now consult Theorem 2.1.1. It may be seen that the requirements that B must satisfy in order to be a Boolean algebra are precisely the first 13 properties mentioned in the theorem, together with one property of the universal set and one of the empty set. The theorem goes on to mention another property of the universal set, one of the empty set and two absorption laws. These are not required in our 'abstract' definition since they follow from the other laws. Indeed, our list of requirements above is itself redundant: a shorter list of axioms is implicit in Exercise 4.4.16.

It also follows from Theorem 3.1.1 that logical equivalence classes of propositional terms form Boolean algebras.

The period from the mid-nineteenth century to the early twentieth century saw the spectacular rise of abstract algebra. In that period of about a hundred years, the meaning of 'algebra' to mathematicians was completely transformed.

Still, in the early nineteenth century 'algebra' meant the algebra of the integers, the rationals, the real numbers and the complex numbers (the last still having dubious status to many mathematicians of the time). Moreover, the rules

of algebra, such as $(a + b) + c = a + (b + c)$ and $ab = ba$, were regarded as fixed and given. The suggestion that there might be 'algebraic systems' obeying different laws would have been almost unintelligible at the time.

Nevertheless, the by then well established use of symbolic notation (as opposed to the earlier ways of presenting arguments, which used far more words) could not help but impress on mathematicians that many elementary arguments involved nothing more than manipulating symbols according to certain rules, and that the rules could be extracted and stated as axioms. For instance, it was noted that such arguments performed with the real numbers in mind also yielded results which were true of the complex numbers. At the time, this was regarded as somewhat puzzling, whereas we would now say that it is because both are fields (of characteristic zero, see Exercise 4.4.11 at the end of the section).

Peacock separated out, to some extent, the abstract algebraic content of such manipulations. But the actual laws were those applicable to the real and the complex numbers: in particular, the commutativity of multiplication was seen as necessary. Gregory and De Morgan continued this work and De Morgan further separated manipulations with symbols from their possible interpretations in particular algebraic systems. Still, the axioms were essentially those for a commutative integral domain. Indeed, the position of many mathematicians was almost that these laws were in some sense universal and necessary axioms, their form deriving simply from the manipulation of symbols.

But this point of view became untenable after Hamilton's development of the quaternions. Hamilton developed the point, between 1829 and 1837, that the meaning of the '+' appearing in the expression of a complex number such as $2 + 3\mathrm{i}$ is quite distinct from the meaning of the symbol '+' in an expression such as $2 + 5$: one could as well write (a, b) for the complex number $a + b\mathrm{i}$ and give the rules for addition and multiplication in terms of these ordered pairs. Thus complex numbers were pairs of real numbers, or 'two-dimensional' numbers, with an appropriately defined addition and multiplication. Because complex numbers could be used to represent forces (for instance) in the plane, there was interest in finding 'three-dimensional' numbers which could be used to represent forces and the like in 3-space. In fact, Gauss had already considered this problem and had, around 1819, come up with an algebra (in which the multiplication was not commutative). But the algebra turned out to be unsuitable for the representation of forces and, as with much of his work, he did not publish it, so that it remained unknown until the publication of his diaries in the later part of the century.

Hamilton searched for many years for such 'three-dimensional' numbers. On the 16th of October 1843, Hamilton and his wife were walking into Dublin

by the Royal Canal. Hamilton had been thinking over the problem of 'three-dimensional numbers' and was already quite close to the solution. Then, in a flash of inspiration, he saw precisely how the numbers had to multiply. (Hamilton later claimed that he scratched the formulae for the multiplication into the stone of Brougham Bridge.) Hamilton had abandoned two preconceptions: that the answer was a three-dimensional algebra, in fact four dimensions were needed, and that the multiplication would be commutative – you can see from the group table (Example 5 of Section 4.3.1) that the quaternions have a non-commutative multiplication. Actually, Grassmann in 1844 published somewhat related ideas but his work was couched in rather obscure terms and this lessened its immediate influence.

Despite Hamilton's hopes, the use of quaternions to represent forces and other physical quantities in 3-space was not generally adopted by physicists: a formalism (essentially vector analysis), due to Gibbs and based on Grassmann's ideas, was eventually preferred. But the effect on mathematics was profound. For this was an algebra with properties very different from the real and complex numbers.

Somewhat later, Boole's development of the algebras we now term Boolean algebras (see above) provided other examples of new types of algebraic systems. Also, the development of matrix algebra and, more particularly, its recognition as a type of algebraic system (by Cayley and B. and C.S. Peirce) provided the, probably more familiar, example of the algebra of $n \times n$ matrices (with, say, real coefficients).

The effect of all this was to free algebra from the presuppositions which had limited its domain of applicability. In the later part of the century many algebras and kinds of algebras were found, and the shift towards abstract algebra – defining algebras in terms of the conditions which they must satisfy rather than in terms of some particular structure – was well under way.

Exercises 4.4

1. Which of the following sets are semigroups under the given operations:
   (i) the set $\mathbb{Z}$ under the operation $a * b = s$, where s is the smaller of a and b;
   (ii) the set of positive integers under the operation $a * b = d$ where d is the gcd of a and b;
   (iii) the set $P(X)$ of all subsets of a set X under the operation of intersection;
   (iv) the set $\mathbb{R}$ under the operation $x * y = x^2 + y^2$;
   (v) the set $\mathbb{R}$ under the operation $x * y = (x + y)/2$.

2. Prove the characterisation of surjections stated at the end of Example 3 (on p. 186).
3. Which of the following are rings? For those which are, decide whether they have identity elements, whether they are commutative and whether they have zero-divisors.
   (i) The set of all 3×3 upper-triangular matrices with real entries, under the usual matrix addition and multiplication.
   (ii) The set of all upper-triangular real matrices with 1s along the diagonal (the form is shown), with the usual operations.
   $$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}$$
   (iii) The set, $P(X)$, of all subsets of a set X, with intersection as the multiplication and with union as the addition.
   (iv) The set, $P(X)$, of all subsets of a set X, with intersection as the multiplication and with symmetric difference as the addition: the **symmetric difference**, $X \Delta Y$, of two sets X and Y is defined to be $X \Delta Y = (X \backslash Y) \Delta (Y \backslash X)$ (see Exercise 2.1.4).
   (v) The set of all integer multiples of 5, with the usual addition and multiplication.
   (vi) The set, $[4]_{24}\mathbb{Z}_{24}$, of all multiples, $[4]_{24}[a]_{24}$, of elements of $\mathbb{Z}_{24}$ by $[4]_{24}$, equipped with the usual addition and multiplication of congruence classes.
   (vii) As (vi), but with $[8]_{24}\mathbb{Z}_{24}$ in place of $[4]_{24}\mathbb{Z}_{24}$.
   (viii) The set of all rational numbers of the form m/n with n odd, under the usual addition and multiplication.
4. Use the axioms for a ring to prove the following facts.
   For all x, y, z in a ring,
   (i) $x(-y) = -xy = -x(y)$,
   (ii) $-1 \cdot x = -x$,
   (iii) $(x - y)z = xz - yz$ (where $a - b$ means, as usual, $a + (-b)$).
5. Show that if the ring R has no zero-divisors, and if a, b and $c \neq 0$ are elements of R such that $ac = bc$, then $a = b$.
6. Show that if x is an idempotent element of the ring R (that is, $x^2 = x$) then $1 - x$ also is idempotent.
7. Find an example of a ring R and elements x, y in R such that $(x + y)^2 \neq x^2 + 2xy + y^2$.
8. Find an example of a ring R and non-zero elements x, y in R such that $(x + y)^2 = x^2 + y^2$.

9. Which of the following are vector spaces over the given field?
   (i) The set of all 2×2 matrices over $\mathbb{R}$, with the usual addition, and multiplication given by
   $$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}.$$
   (ii) The set of all 2×2 matrices over $\mathbb{R}$, with the usual addition, and multiplication given by
   $$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda a & b \\ c & \lambda d \end{pmatrix}.$$
   (iii) The set of all 2×2 matrices over $\mathbb{R}$, with the usual addition, and multiplication given by
   $$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda^{-1} a & b \\ c & \lambda^{-1} d \end{pmatrix}$$
   for $\lambda \neq 0$ and
   $$0 \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$
10. Deduce the following consequences of the axioms for a vector space. For any vectors x, y and scalar λ,
   (i) $0 \cdot x = 0$ ('0' here is the scalar zero).
   (ii) $\lambda \cdot 0 = 0$ (here, '0' means the zero vector),
   (iii) $\lambda(x - y) = \lambda x - \lambda y$,
   (iv) $-1 \cdot x = -x$.
11. Let F be any field. The **characteristic** of F is defined to be the least positive integer n such that $1 + 1 + \cdots + 1 = 0$ (n 1s). If there is no such n then the characteristic of F is said to be 0. Show that if the characteristic of F is not zero then it must be a prime number.
12. Suppose that R is an integral domain. Let X be the set of all pairs (r, s) of elements of R with $s \neq 0$. Define a relation $\sim$ on X by $(r, s) \sim (t, u)$ iff $ru = st$. Show that $\sim$ is an equivalence relation. Let Q be the set of all equivalence classes $[(r, s)]$ of $\sim$. Define an addition and multiplication on Q by $[(r, s)] + [(t, u)] = [(ru + st, su)]$ and $[(r, s)] \times [(t, u)] = [(rt, su)]$. Show that these operations are well defined (i.e. do not depend on the chosen representatives of the equivalence classes). Show that Q is a commutative ring under these operations. Show that every non-zero element of Q has an inverse and hence that Q is a field.

Define a function $f\colon R \to Q$ by sending $r \in R$ to the equivalence class of $(r, 1)$. Show that this is an injection, that $f(r + t) = f(r) + f(t)$, that $f(rt) = f(r) \times f(t)$.
Thus there is a copy of the ring R sitting inside the field Q. Finally, show that every element of Q has the form $f(r) \cdot f(s)^{-1}$ for some $r, s \in R$ with $s \neq 0$. That is, Q is essentially the field consisting of all fractions formed from elements of R: Q is called the **field of fractions** (or **quotient field**) of R. You can check that if the initial ring R is the ring $\mathbb{Z}$ of integers then Q can be thought of as the field $\mathbb{Q}$ of rational numbers; if we start with the integral domain $R = \mathbb{Z}[\sqrt{2}]$ then we end up with a copy of the field $\mathbb{Q}[\sqrt{2}]$. [Hint: think of the pair (r, s) as being a 'fraction' r/s.]

13. Let F be the following set of matrices with entries in $\mathbb{Z}_2$, under matrix addition and multiplication (we write '0' for $[0]_2$, '1' for $[1]_2$):

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Show that F is a field and is also a $\mathbb{Z}_2$-algebra, where the elements λ of $\mathbb{Z}_2$ act by

$$\lambda \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}.$$

You should check that this is a field of characteristic 2 in the sense of Exercise 4.4.11.

14. Combine Example 4 of Section 4.3.3 with Example 3 on p. 192 to realise the ring $\mathbb{H}$ of quaternions as an $\mathbb{R}$-algebra of 4×4 matrices with real entries.
15. Let B be a Boolean algebra. Show that
    (1) $a \wedge 0 = 0$,
    (2) $a \vee (a \wedge b) = a$,
    (3) $a \vee (\neg a \wedge b) = a \vee b$.
16. (a) Let $B = (B; \wedge, \vee, \neg, 0, 1)$ be a Boolean algebra. Define new operations, '$+$' and '$\cdot$', on the set B by

$$a \cdot b = a \wedge b,$$
$$a + b = (a \vee b) \wedge \neg(a \wedge b).$$

(Observe that, in the context of algebras of sets, $a + b$ is the symmetric difference of a and b, see Exercise 4.4.3 (iv) above.)
    (i) Show that, with these operations, B is a commutative ring.
    (ii) Identify the zero element of this ring and the identity element.

(iii) Show that every element is idempotent and that for every $a \in B$ $a + a$ is the zero element.

(b) Suppose that $(R; +, \cdot, 0, 1)$ is a commutative ring in which every element a satisfies $a^2 = a$ and $a + a = 0$ (such a ring is termed a **Boolean ring**). Define operations '$\wedge$' and '$\vee$' on R by

$$a \wedge b = a \cdot b$$
$$a \vee b = a + b + a \cdot b$$

Show that $(R, \wedge, \vee, 0, 1)$ is a Boolean algebra.

Show also that if we start with a Boolean algebra, produce the ring as in (a), and then go back to a Boolean algebra as in (b), then we recover the Boolean algebra with which we began. Similarly the process (b) applied to a Boolean ring, followed by (a), recovers the original ring.

Thus one may say that Boolean algebras and Boolean rings are equivalent concepts.

17. Suppose that $(B, \wedge, \vee, \neg)$ is a Boolean algebra. Define a relation $\leq$ on B by $a \leq b$ iff $a = a \wedge b$. Show that $a = a \wedge b$ iff $b = a \vee b$. Prove that $\leq$ is a partial order on the set B and that $0 \leq a \leq 1$ for all $a \in B$. What is the relation $\leq$ in the case that B is a Boolean algebra $(B, \cap, \cup, ^c)$ of sets?

In fact (you may try to show this as a further exercise) the Boolean operations $\wedge$, $\vee$ and $\neg$ may be defined in terms of this partial order and hence Boolean algebras may be regarded as certain kinds of partially ordered sets.

Summary of Chapter 4

In Section 4.1 we discussed permutations, including their cycle decomposition. The order and sign of a permutation were considered in the next section. In Section 4.3 we introduced the definition of a group and gave many examples of this concept. In the fourth section we gave a brief discussion of various other algebraic structures which appear in this book.

5 Group theory and error-correcting codes

By now we have met many examples of groups. In this chapter, we begin by considering the elementary abstract theory of groups. In the first section we develop the most immediate consequences of the definition of a group and introduce a number of basic concepts, in particular, the notion of a subgroup. Our definitions and proofs are abstract, but are supported by many illustrative examples. The main result in this chapter is Lagrange's Theorem, which is established in Section 5.2. This theorem says that the number of elements in a subgroup of a finite group divides the number of elements in the whole group. The result has many consequences and provides another proof of the theorems of Fermat and Euler which we proved in Chapter 1. In the third section we define what it means for two groups to be isomorphic: to have the same abstract form. Then, after describing a way of building new groups from old, we move on to describe, up to isomorphism, all groups with up to eight elements. The final section of the chapter gives an application of some of the ideas we have developed, to error-detecting and error-correcting codes.

5.1 Preliminaries

We introduced the idea of an abstract group in Section 4.3 and then gave many examples of groups. In this section we will prove a number of results which hold true for all these examples. For instance, we do not, every time we wish to refer to the inverse of an element of a group, prove that the inverse of that element is unique. Rather, we proved once and for all that if a is an element of a group then there is a unique inverse for a (Theorem 4.3.1). Then the result applies to any particular element in any particular group. This is one of the main advantages of working in the abstract: we may deduce a result once and for all without having to prove it again and again in particular cases. In this section,

we will consider some elementary deductions from the definition of a group. Then the central concept of a subgroup is introduced.

We start by making some observations which may, at first sight, seem of little significance.

We are going to use the four group axioms to make further deductions. In doing this, we will only employ the usual rules of reasoning together with our four axioms. However, we need to be clear as to what the rules of reasoning say in this context. Suppose we have an equation such as $a = b$ for certain expressions, a and b in a group G. A correct deduction from this would be achieved by performing the same operation to both sides of our equation $a = b$. Care is needed when we carry this out. If, for example, we multiply both sides by a group element x, we must either multiply both sides by x on the right (to obtain $ax = bx$), or by x on the left (to obtain $xa = xb$). It would be incorrect to deduce that $ax = xb$. The underlying reason for this is that we cannot assume that our group is Abelian.

Example Suppose that in a given group, G, we know that the given elements g and h commute, so $hg = gh$. Then we can multiply both sides of this equation on the right by g^{-1} to obtain $g^{-1}hg = g^{-1}gh = (g^{-1}g)h = eh = h$ (we have made use of group axioms (G2), (G3) and (G4) from p. 170 in these equations). In fact this argument could be reversed: if $g^{-1}hg = h$ then $hg = gh$ (multiply by g on the right). A further deduction from $h = g^{-1}hg$ could be made by multiplying on the right again by h^{-1} to obtain $e = h^{-1}h = h^{-1}g^{-1}hg$ with this last argument again being reversible. Now work from the left by multiplying by g^{-1}, to obtain (after simplification) $g^{-1} = h^{-1}g^{-1}h$. Once again this last argument could be reversed. As a final step, multiply through by h^{-1} to obtain $g^{-1}h^{-1} = h^{-1}g^{-1}$ and we have shown that $hg = gh$ if and only if $h^{-1}g^{-1} = g^{-1}h^{-1}$. That is, two elements commute if and only if their inverses commute.

Our first result, Theorem 5.1.1 below, concerns solvability of certain equations in groups. Since it is our first really abstract result, it may be appropriate (in the spirit of Chapter 1) to give a more detailed commentary on both its statement and its proof. For the statement, note that this is a result about an arbitrary group. Since we are now operating at a higher level of abstraction than in Chapter 1, there are now two ways in which we can illustrate the statement of Theorem 5.1.1. In the first place, we could apply the statement in the context where we have made a particular choice for the group G mentioned in the statement. Alternatively, we could continue to keep our group G as a completely general one, but choose specific values for the elements a or b (such as $a = e$ or $b = a$). In the second sentence, we are concerned with solutions of equations

like $a = bx$. Two points are worth making here. One is a point easily missed in a first reading: we say x is an element of G! Although it may seem clear that if $ax = b$, then (multiplying on the left by a^{-1} and simplifying), we can find a value for x, we must show that x is an element of our group G. The second point is that we claim this solution is unique: we do not just need to find some solution, we must show that it is the only one. Note also that we discuss solutions to two equations $a = bx$ and $a = yb$. We use different letters x and y to emphasise that in a general group the solutions of these two equations could well be different elements of G. After the proof, we will state a corollary before making some comments about the method of proof.

Theorem 5.1.1 *Let G be a group and let a and b be elements of G. Then there are unique elements x and y in G such that $a = bx$ and $a = yb$.*

Proof We first consider the equation $a = bx$ and show that this equation does indeed have a solution. Then we show that there is only one solution. To see that a solution exists, we take x to be $b^{-1}a$. Since $b^{-1} \in G$ and G is closed under products this is an element of G. Then

$$\begin{aligned} bx &= b(b^{-1}a) \\ &= (bb^{-1})a \text{ by associativity (G2) in the definition of a group,} \\ &= ea \text{ by existence of inverse (G4),} \\ &= a \text{ by existence of identity (G3),} \end{aligned}$$

so a solution exists. If c and d both are solutions of $a = bx$, so $a = bc = bd$, then multiply both sides of the equation $bc = bd$ on the left by the inverse of b to obtain

$$\begin{aligned} b^{-1}(bc) &= b^{-1}(bd) \text{ hence} \\ (b^{-1}b)c &= (b^{-1}b)d \text{ by associativity, and so} \\ ec &= ed \text{ by (G4), giving} \\ c &= d \end{aligned}$$

as required. The proof for the equation $a = yb$ is similar and is left as an exercise for the reader. □

Remark The theorem allows us to 'cancel' in a group, provided we do this 'on the same side'. If g, h and b are elements of a group G and $bg = bh$, then g and h must be equal. Similarly, if $gb = hb$ we can deduce $g = h$. This is why no element of a group occurs twice in any row (or column) of a group multiplication table (as we saw in Section 4.3).

Example 1 Let a, b, c be elements of a group G. Find a group element x such that $xaba^{-1} = c$.

To do this, remember to multiply consistently (on the right in this case) and also take the argument a step at a time. First multiply (on the right) by a to obtain

$$ca = (xaba^{-1})a = ((xab)(a^{-1}))a = (xab)(a^{-1}a) = xabe = xab.$$

Now multiply by b^{-1} on the right to obtain (after simplification) $xa = cab^{-1}$. Finally, multiply by a^{-1} on the right to obtain our solution $x = cab^{-1}a^{-1}$. Notice that there is no way to simplify the expression $cab^{-1}a^{-1}$ in general.

We can also solve equations using 'mixed' (right and left) terms but again care is required.

Example 2 Let a, b, c be elements of a group G. Find a group element x such that $axb = b^{-1}c$.

Multiply by b^{-1} on the right to obtain $ax = b^{-1}cb$. Now multiply by a^{-1} on the left to get that $x = a^{-1}b^{-1}cb$.

Corollary 5.1.2 *Let G be a group. Then the identity element of G is unique, inverses are unique, $(a^{-1})^{-1}$ is a and $(ab)^{-1}$ is $b^{-1}a^{-1}$.*

Proof The first two parts have already been established in Theorem 4.3.1 (they may also be viewed as consequences of Theorem 5.1.1: take $a = b$ to deduce that the identity is unique, and take $a = e$ to deduce that inverses are unique). The fact that $(a^{-1})^{-1} = a$ follows since $(a^{-1})^{-1}$ and a both are solutions of $a^{-1}x = e$. Also $(ab)^{-1} = b^{-1}a^{-1}$ since both solve the equation $(ab)x = e$. □

Comment on the proof of Theorem 5.1.1 As we have already noted, there are two ways to specialise a proof in order to come to terms with its abstractness. We could see what the proof would say in a particular circumstance we already know. Thus, in Theorem, 5.1.1, we could choose $a = b$, and then see why (as indicated in the Corollary 5.1.2) the proof provides us with a proof that inverses are unique in a group. Alternatively, we could focus on a specific group we feel familiar with, and try out the detailed steps of the argument in the context of that group.

Once we start on the details of the proof, note that most lines of the proof are, not just an equation, but include some comment on how the equation was obtained. These steps are all either uses of elementary logic (such as performing the same operation to both sides of a known equation, or substituting a known equality into a given equation), or are justified by using one of the four group

axioms. It is possible that the reader might feel quite confident about the steps making up the proof, but wonder how it is possible to prove something for oneself without a model to imitate. The first requirement is really to understand every detail of an argument like this. Then try to see the proof as a whole and try to understand how the informal idea(s) behind the proof can, step-by-step, be transformed into a rigorous, formal proof.

We have already met the idea of taking powers of elements in several contexts, so the following definition should come as no surprise.

Definition Let g be an element of a group G. The positive **powers** of g are defined inductively by setting $g^1 = g$ and $g^{k+1} = gg^k$. We can also define zero and negative powers by putting $g^0 = e$ and $g^{-k} = (g^{-1})^k$ for $k > 0$.

Note that this definition implies, for example that g^2 means $g \cdot g$. This may again seem a trivial point, but if g is itself a product of two other group elements, say $g = xy$, then g^2 will mean $(xy)(xy)$. This is not the same as x^2y^2 in general.

The next result gives the index laws for group elements.

Theorem 5.1.3 *Let G be a group and let g and h be elements of G. For any integers r and s we have*

(i) $g^r g^s = g^{r+s}$,
(ii) $(g^r)^s = g^{rs}$,
(iii) $g^{-r} = (g^r)^{-1} = (g^{-1})^r$, *and*
(iv) *if* $gh = hg$, *then* $(gh)^r = g^r h^r$.

Proof (iv) For non-negative integers, the proof of part (iv) is just like the proof of Theorem 4.2.1 (iv). If r is negative, say $r = -k$ for some positive integer k, then we have

$$\begin{aligned} g^r h^r &= (g^{-1})^k (h^{-1})^k \text{ by definition} \\ &= (g^{-1}h^{-1})^k \text{ since } k \text{ is positive} \\ &= ((hg)^{-1})^k \text{ by Corollary 5.1.2 and since, as we verified above,} \\ &\qquad\qquad gh = hg \text{ implies } h^{-1}g^{-1} = g^{-1}h^{-1} \\ &= ((gh)^{-1})^k \text{ by assumption} \\ &= (gh)^{-k} \text{ by definition} \\ &= (gh)^r \text{ as required.} \end{aligned}$$

(iii) Apply part (iv) with $h = g^{-1}$ to get $e = e^r = (gg^{-1})^r = g^r(g^{-1})^r$ for every integer r. Therefore, $(g^r)^{-1} = (g^{-1})^r$ and the latter, by definition, is g^{-r}.

(i) The case where both r and s are non-negative is proved just as in 4.2.1 (i). So, in treating the other cases, we may suppose that at least one of r, s is negative. We split this into three further cases: (a) the case when $r + s > 0$, (b) the case when $r + s = 0$, and (c) the case when $r + s < 0$.

(a) If $r + s > 0$ then at least one of r, s must be strictly greater than 0: say $r > 0$ (the argument supposing that $s > 0$ is similar). So, by assumption, $s < 0$ and hence $-s > 0$. Then, by the case where both integers are positive we have

$$g^{r+s}g^{-s} = g^{r+s-s} = g^r$$

so, multiplying on the right by g^s, we obtain

$$g^{r+s} = g^r g^s,$$

as required.

(b) When $r + s = 0$, then $s = -r$ so $g^r g^s = g^r g^{-r} = e$ by part (iii). But, by definition, $e = g^0 = g^{r+s}$.

(c) When $r + s < 0$, then $-r + (-s) > 0$. So, by the case where both integers are positive, we have

$$g^{-(r+s)} = g^{-s+(-r)} = g^{-s}g^{-r}.$$

By part (iii), the inverse of $g^{-(r+s)}$ is g^{r+s}: by 5.1.2 and part (iii), the inverse of $g^{-s}g^{-r}$ is $g^r g^s$. So we conclude $g^{r+s} = g^r g^s$.

(ii) The proof of this part follows by induction from part (i). □

It is a consequence of the first part of the above result that if g is an element of a group G and r, s are integers then

$$g^r g^s = g^{r+s} = g^{s+r} = g^s g^r.$$

That is, the powers of an element g commute with each other.

Next we define the order of an element of a group, another idea which should be familiar from Chapters 1 and 4.

Definition An element g of a group G is said to have **infinite order** if there is no positive integer n for which $g^n = e$. Otherwise, the **order** of g is the smallest positive integer such that $g^n = e$.

The following result is proved in precisely the same way as is Theorem 4.2.3.

Theorem 5.1.4 *Let g be an element of a group G and suppose that g has finite order n. Then $g^r = g^s$ if and only if r is congruent to s modulo n.*

Example 1 The order of a permutation, as defined in Section 4.2, is, of course, a special case of the general definition above. As we saw in Section 4.2, the order of a permutation π in the group $S(n)$ may easily be calculated in terms of the expression of π as a product of disjoint cycles. In a general group G, however, there will be no easy way to predict the order of an element.

Example 2 If G is a finite group then every element must have finite order (the proof is just like that for Theorem 4.2.2). So, to find elements of infinite order we must go to infinite groups. Let $\mathrm{GL}(2, \mathbb{R})$ be the group of invertible 2×2 matrices with real entries. The matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

has infinite order since (as may be proved by mathematical induction) A^n is the matrix

$$A^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}.$$

If, in this example, we were to replace the field $\mathbb{R}$ by the field $\mathbb{Z}_p$ for some prime p, so we would consider 2×2 invertible matrices with entries in $\mathbb{Z}_p$, then the matrix

$$A = \begin{pmatrix} [1]_p & [1]_p \\ [0]_p & [1]_p \end{pmatrix}$$

would have order p:

$$A^p = \begin{pmatrix} [1]_p & [p]_p \\ [0]_p & [1]_p \end{pmatrix} = \begin{pmatrix} [1]_p & [0]_p \\ [0]_p & [1]_p \end{pmatrix} = I.$$

We now come to one of the key ideas in elementary group theory.

Definition A non-empty subset H of a group G (more precisely, of $(G, *)$) is a **subgroup** of G if H is itself a group under the same operation $(*)$ as that of G (or, more precisely, under the operation of G restricted to H).

In particular, it must be that a subgroup H of a group G contains the identity element e of G (for if $f \in H$ acts as an identity in H then, working in G, from $ff = f = ef$, we deduce $f = e$) and the inverse of any element of H lies in H and is just its inverse in G.

In order to check whether or not a given subset of G is a subgroup, it would appear that we need to check the four group axioms. However, the next result shows that it is sufficient to check rather less. After establishing this, we will consider some standard examples of subgroups.

Theorem 5.1.5 *The following conditions on a non-empty subset H of a group G are equivalent.*

(i) *H is a subgroup of G.*
(ii) *H satisfies the following two conditions:*
  (a) *if h is in H then h^{-1} is in H; and*
  (b) *if h and k are in H then hk is in H.*
(iii) *If h and k are in H then hk^{-1} is in H.*

Proof It has to be shown that the three conditions are equivalent. What we will do is show that (i) implies (ii), (ii) implies (iii), (iii) implies (i). It then follows, for example, that (i) implies (iii) and indeed that the three conditions are equivalent.

That (i) implies (ii) follows directly, since the conditions in (ii) are two of the group axioms. It is also easy to see that (ii) implies (iii). For if h and k are in H then h and k^{-1} are in H (H is closed under taking inverses by (ii)(a)) and then hk^{-1} is in H (by (ii)(b)). So it only remains to show that (iii) implies (i).

We check the four group axioms for H. First note that the associativity axiom holds since if g, h and k are elements of H then certainly g, h and k are elements of G and so $(gh)k = g(hk)$. Next, we show that H contains the identity element of G. To see this take any h in H (this is possible since H is non-empty) and apply (iii) with $h = k$ to obtain that $e = hh^{-1}$ is in H. Now let g be any element of H and apply (iii) with h being e (which we now know to be in H) and k being g to see that (iii) implies that g^{-1} must be in H. Finally we check the closure axiom for H. Given x and y in H, we have just seen that y^{-1} must also be in H: applying (iii) with $h = x$ and $k = y^{-1}$ gives $xy \in H$ (since $(y^{-1})^{-1} = y$). □

Example 1 Let H be the set of even integers, considered as a subset of $G = (\mathbb{Z}, +)$.

In order to apply Theorem 5.1.5 we need first to check that the subset we are considering is not the empty set. In this case, there seems very little to say: of course there are even numbers! However, in general we might need to be more careful. A good habit to acquire is to check if the identity element of the whole group G is in the subset H (if so, this proves that H is non-empty, but we have already seen that every subgroup of G must contain the identity element of G, so we have not done unnecessary work). In this case, the identity element of G is the number 0 (since G is a group under addition). Also 0 is even because it is divisible by 2 (since $0 = 0 \cdot 2$).

We now check conditions (a) and (b) of (ii): they follow because the sum of two even integers is an even integer and the negative of an even integer is an even integer. We have shown that H is a subgroup of G

Example 2 The set $(\mathbb{Z}, +)$ is itself a subgroup of $(\mathbb{R}, +)$ which is in turn a subgroup of $(\mathbb{C}, +)$.

Example 3 Let H be the set, $A(n)$, of even permutation in the group $S(n)$ under composition of permutations.

The identity element of $S(n)$ is an even permutation, and so is in $A(n)$. By Theorems 4.2.8 and 4.2.9, the inverse of an even permutation is an even permutation and the sign of a product of two permutations is the product of the signs. It follows that $A(n)$ is a subgroup of $S(n)$.

Example 4 Next let H be the set of invertible diagonal 2×2 matrices, considered as a subset of the group, $\mathrm{GL}(n, \mathbb{R})$, of all invertible 2×2 matrices under matrix multiplication.

Again it is easy to check that the identity element for G (the identity matrix) is in H (because the identity matrix is diagonal). Since the product of two diagonal matrices is a diagonal matrix and the inverse of a diagonal matrix is also a diagonal matrix, we deduce that H is a subgroup of G.

Example 5 Let H be the set of all $n \times n$ matrices with determinant 1, considered as a subset of $\mathrm{GL}(n, \mathbb{R})$.

Again, we use part (ii) of the above theorem and check conditions (a) and (b). Note that the identity element of G has determinant 1, so is in H. If A, B are matrices with determinant 1, then AB also has determinant 1. The determinant of the inverse of an invertible matrix A is equal to 1 over the determinant of A, so if A has determinant 1, the determinant of A^{-1} is also equal to 1. Thus H is a subgroup, which is usually denoted $\mathrm{SL}(n, \mathbb{R})$.

Remarks (1) The advantage of checking for a subgroup in this systematic way is that we will immediately detect (if any of our checks fails) if a given subset is not a subgroup of G.

(2) Every group has obvious subgroups, namely the group G itself (any other subgroup is said to be **proper**) and the **trivial** or **identity** subgroup $\{e\}$ containing the identity element only. These will be distinct provided G has more than one element.

As a simple application of this result we show the following.

Theorem 5.1.6 *Let G be a group and let H and K be subgroups of G. Then the intersection $H \cap K$ is a subgroup of G.*

Proof Note that $H \cap K$ is non-empty since both H and K contain e. We show that $H \cap K$ satisfies condition (iii) of Theorem 5.1.5. Take x and y in $H \cap K$: then x and y are both in H and so, since H is a subgroup, xy^{-1} is in H. Similarly, since x and y are both in K and K is a subgroup, xy^{-1} is in K. Hence xy^{-1} is in $H \cap K$. So, by Theorem 5.1.5, $H \cap K$ is a subgroup of G. □

A good source of subgroups is provided by the following.

Theorem 5.1.7 *Let G be a group and let g be an element of G. The set $\langle g \rangle = \{g^n : n \in \mathbb{Z}\}$ of all distinct powers of g is a subgroup, known as the subgroup* **generated by** *g. It has n elements if g has order n and it is infinite if g has infinite order.*

Proof To see that $\langle g \rangle$ is a subgroup, note first that it is non-empty since it contains g. If h and k are in $\langle g \rangle$ then h is g^i and k is g^j for some integers i and j, and so

$$\begin{aligned} hk^{-1} &= g^i(g^j)^{-1} \\ &= g^i g^{-j} \\ &= g^{i-j} \text{ by Theorem 5.1.3.} \end{aligned}$$

Thus hk^{-1} is in $\langle g \rangle$ and so, by Theorem 5.1.5, $\langle g \rangle$ is a subgroup of G as required.

If g has infinite order, then the positive powers $g, g^2, \ldots, g^n, \ldots$ are all distinct (see the proof of Theorem 4.2.2) so $\langle g \rangle$ is an infinite group. If g has order n, then Theorem 5.1.4 shows that there are n distinct powers of g and hence that $\langle g \rangle$ has exactly n elements. □

Definition A group of the above type, that is, of the form $\langle g \rangle$ for some element g in it, is said to be **cyclic**, **generated by** g.

Remark It follows from Theorem 5.1.3 that a cyclic group is Abelian.

Example 1 To find all the cyclic subgroups of $S(3)$, we may take each element of the group $S(3)$ in turn and compute the cyclic subgroup which it generates. In this way, we obtain a complete list (with repetitions) as follows:

$$\begin{aligned} \langle \text{id} \rangle &= \{\text{id}\}; \langle (12) \rangle = \{\text{id}, (12)\}; \\ \langle (13) \rangle &= \{\text{id}, (13)\}; \langle (23) \rangle = \{\text{id}, (23)\}; \\ \langle (123) \rangle &= \{\text{id}, (123), (132)\} = \langle (132) \rangle. \end{aligned}$$

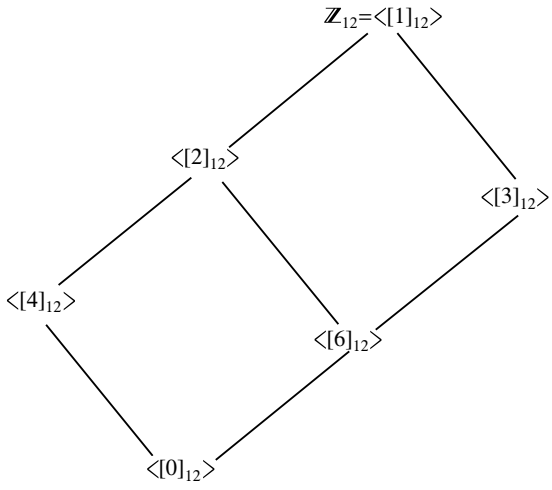


Fig. 5.1

Since we have listed all the cyclic subgroups in $S(3)$ and since $S(3)$ itself is not on the list, it follows that the group $S(3)$ is not cyclic.

Note that, in principle, we can find all the cyclic subgroups of a given finite group by taking each element of the group in turn and computing the cyclic subgroup it generates, then deleting any duplicates.

Example 2 As we saw above, the subgroup of $\mathrm{GL}(2, \mathbb{R})$ generated by

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

is an infinite cyclic group.

Example 3 In the additive group $(\mathbb{Z}, +)$, the cyclic subgroup generated by the integer 1 consists of all multiples (not powers, because the group operation is addition!) of 1 and so is the whole group. The subgroup $\langle 2\rangle$ is the set of even integers. The subgroup $\langle 3\rangle$ consists of all integer multiples of 3. The intersection is easily computed: $\langle 2\rangle \cap \langle 3\rangle = \langle 6\rangle$.

Example 4 The group $(\mathbb{Z}_{12}, +)$ is itself cyclic, as are all its subgroups. The set of all its subgroups forms a partially ordered set under inclusion. The Hasse diagram of this set is shown in Fig. 5.1.

The reader may be familiar with vector spaces, where the change from structures generated by one element to those generated by two is not very great.

The situation for groups is very different. A group generated by two elements may well be immensely complicated. In Section 4.3.4 we saw some of the simpler examples: the dihedral groups (groups of symmetries of regular n-sided polygons) can be generated by two elements. These elements are the reflection R in the perpendicular bisector of any one of the sides and the rotation ρ, through $2\pi/n$ radians. Thus $R^2 = e = \rho^n$ and there is also the relation $\rho^{n-1}R = R\rho$. The resulting group has $2n$ elements which can all be written in the form $\rho^i R^j$ where j is 0 or 1 and $0 \leq i < n$.

We also saw in Section 4.2 that the symmetric group $S(n)$ can be generated by its transpositions. In fact $S(n)$ can be generated by the $n-1$ transpositions $(1\ 2), (2\ 3), \ldots, (n-1\ n)$: for every element of $S(n)$ can be written as a product of these. (This was set as Exercise 4.2.10.) However, the relations between these generators are much more complicated than in the case of the dihedral groups.

Exercises 5.1

1. Prove that for any elements a and b of a group G, $ab = ba$ if and only if $(ab)^{-1} = a^{-1}b^{-1}$.
2. Let a, b be elements of a group G. Find (in terms of a and b) an expression for the solution x of the equation $axba^{-1} = b$.
3. Take G to be the cyclic group with 12 elements. Find an element g in G such that the equation $x^2 = g$ has no solution.
4. Use Theorem 5.1.1 to decide which of the following subsets of the given groups are subgroups:
   (i) the subset of the symmetries of a square consisting of the rotations;
   (ii) the subset of $(\mathbb{R}, +)$ consisting of $\mathbb{R}\backslash\{0\}$ under multiplication;
   (iii) the subset $\{\text{id}, (1\ 2), (1\ 3), (2\ 3)\}$ of $S(3)$;
   (iv) the subset of $(\mathrm{GL}(3, \mathbb{R}), \cdot)$ consisting of matrices of the form
   $$\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}.$$
5. Give an example of a group G and elements a, b, and c in G, such that a is different from b but $ac = cb$.
6. Let G be a cyclic group generated by x. Note that for any positive integer k, the set $\langle x^k \rangle$ is a subgroup of G. If x has finite order 12, consider the possible values for k (from 0 to 11) and, in each case, show that $\langle x^k \rangle$ is generated by x^d where d is the greatest common divisor of k and n. Deduce that $\langle x^k \rangle$ has $12/d$ elements.

7. Let G be a cyclic group generated by x with x of order greater than 1. Let H be a subgroup of G with H neither G nor $\{e\}$. Let $m \geq 1$ be minimal such that x^m is in H. Use the division algorithm to show that x^m generates H and deduce that every subgroup of a cyclic group is cyclic.
8. Let g and x be elements of a group G. Show that for all positive integers k,
$$(g^{-1}xg)^k = g^{-1}x^k g.$$
Deduce that x has order 3 if and only if (for all $g \in G$) $g^{-1}xg$ has order 3. Show that the same is true for any integer $n \geq 1$ in place of 3.
9. Let G be any group and define the relation of conjugacy on G by aRb if and only if there exists $g \in G$ such that $b = g^{-1}ag$. Show that this is an equivalence relation on G.
10. Find $a \in G_{23}$, the group of invertible congruence classes modulo 23, such that every element of G_{23} is a power of a: that is, show that G_{23} is a cyclic group by finding a generator for it. Similarly show that G_{26} is cyclic by finding a generator for it. Is every group of the form G_n cyclic?

5.2 Cosets and Lagrange's Theorem

If H is a subgroup of a group G then G breaks up into 'translates', or *cosets*, of H. This notion of a coset is a key concept in group theory and we will make use of cosets in this section by proving Lagrange's Theorem. The remainder of the section is devoted to deriving consequences of this theorem, which may be regarded as the fundamental result of elementary group theory.

Definition Let H be a subgroup of the group G, and let a be any element of G. Define aH to be the set of all elements of G which may be written as ah for some element h in H: $aH = \{ah : h \in H\}$. This is a (**left**) **coset** of H (in G): it is also termed the left coset of a with respect to H. Similarly define the **right coset** $Ha = \{ha : h \in H\}$. We make the convention that the unqualified term 'coset' means 'left coset'.

Notes (1) The subgroup H is a coset of itself, being equal to eH.

(2) The element a is always a member of its coset aH since $a = ae \in aH$ (for e is in H, since H is a subgroup). Similarly, a is a member of the right coset Ha.

(3) If b is in aH then $bH = aH$. To see this suppose that $b = ah$ for some h in H. A typical member of bH has the form bk for some k in H. We have:
$$bk = (ah)k = a(hk).$$

Since H is a subgroup, hk is in H and so we have that bk is in aH. Thus $bH \subseteq aH$.

For the converse, note that from the equation $b = ah$ we may derive $a = bh^{-1}$, and h^{-1} is in H. So we may apply the argument just used to deduce $aH \subseteq bH$. Hence $aH = bH$ as claimed.

It follows that each coset of a given subgroup is determined by any one of the elements in it. Such an element is known as a **representative** for the coset.

(4) Unless a is in H the coset aH is not a subgroup (for if it were a subgroup it would have to contain the identity, so we would have $e = ah$ for some h in H, necessarily $h = a^{-1}$, but then since a^{-1} is in H we would have that $a = (a^{-1})^{-1}$ is a member of H).

(5) There are two 'trivial' cases: if $H = G$ then there is only one coset of H, namely $H = G$ itself; if $H = \{e\}$ then for every a in G the coset aH consists of just a itself.

Example 1 Take $G = (\mathbb{Z}, +)$ and let $n \geq 2$ be a positive integer. Let H be the set of all integer multiples of n: note that H is a subgroup of $\mathbb{Z}$ (proved as in Section 5.1). (We exclude the values $n = 1$ and $n = 0$ since they correspond to the two trivial cases mentioned in Note (5) above.) What are the cosets of H in G? We have already met them! For example H consists of precisely the multiples of n and so is just the congruence class of 0 modulo n. Similarly the coset $1 + H$ (we use additive notation since the operation in G is '+') is none other than the congruence class of 1 modulo n: and in general the coset $k + H$ of k with respect to H is just the congruence class of k modulo n. You may note that in this example right and left cosets coincide: $k + H = H + k$, since the group is Abelian.

Example 2 Take $G = (\mathbb{Z}_6, +)$ and let H be the subgroup with elements 0 and 3 (more precisely $[0]_6$ and $[3]_6$): this set of two elements does form a subgroup, being the subgroup generated by 3. The cosets are as follows:

$$0 + \{0,\ 3\} = \{0,\ 3\} = 3 + \{0,\ 3\};$$
$$1 + \{0,\ 3\} = \{1,\ 4\} = 4 + \{0,\ 3\};$$
$$2 + \{0,\ 3\} = \{2,\ 5\} = 5 + \{0,\ 3\}.$$

Observe that each of the three cosets of H in G has the same number (two) of elements as H (this is a general fact, see Theorem 5.2.2 below). We may also note that each coset has two representatives.

Example 3 Take G to be the symmetric group $S(3)$ with the usual composition of permutations, and let H be the subgroup consisting of the identity element together with the transposition (1 2) (in the terminology of Section 4.1, this is

the subgroup generated by (1 2)). To find the complete list of cosets of H in G, we may consider all sets of the form gH for $g \in G$. In this way we get a list of six cosets of H in G, but these are not all distinct:

$$\begin{aligned}
\mathrm{id} \cdot H &= \mathrm{id} \cdot \{\mathrm{id}, (1\ 2)\} = \{\mathrm{id}, (1\ 2)\};\\
(1\ 2)H &= (1\ 2)\{\mathrm{id}, (1\ 2)\} = \{(1\ 2), (1\ 2)(1\ 2)\} = H;\\
(1\ 3)H &= \{(1\ 3), (1\ 3)(1\ 2)\} = \{(1\ 3), (1\ 2\ 3)\};\\
(2\ 3)H &= \{(2\ 3), (2\ 3)(1\ 2)\} = \{(2\ 3), (1\ 3\ 2)\};\\
(1\ 2\ 3)H &= \{(1\ 2\ 3), (1\ 2\ 3)(1\ 2)\} = \{(1\ 2\ 3), (1\ 3)\};\\
(1\ 3\ 2)H &= \{(1\ 3\ 2), (1\ 3\ 2)(1\ 2)\} = \{(1\ 3\ 2), (2\ 3)\}.
\end{aligned}$$

We see that $\mathrm{id} \cdot H = (1\ 2)H, (1\ 3)H = (1\ 2\ 3)H$ and $(2\ 3)H = (1\ 3\ 2)H$. Again we note that each coset contains two elements and so has two representatives. Notice that the right and left cosets of a given element with respect to H need not coincide:

$$H(1\ 3) = \{(1\ 3), (1\ 3\ 2)\} \neq (1\ 3)H.$$

In fact the right coset $H(1\ 3)$ is not the left coset of any element.

Example 4 Take G to be Euclidean 3-space ($\mathbb{R}^3$) with addition of vectors as the operation. Let H be the xy-plane (defined by the equation $z = 0$). Note that H is a subgroup of G. You should check that the cosets of H in G are the horizontal planes: indeed the coset of a vector $\mathbf{v}$ is just the plane which contains $\mathbf{v}$ and is parallel to H (the horizontal plane containing $\mathbf{v}$).

Next, we see why (as you may have noticed in our examples) if two cosets of a given subgroup H have an element in common, then they are equal. Indeed, in the proof we show that the relation on G defined by

xRy if and only if $y^{-1}x \in H$

is an equivalence relation which partitions G into the distinct cosets of H.

Theorem 5.2.1 *Let H be a subgroup of the group G and let a, b be elements of G. Then either $aH = bH$ or $aH \cap bH = \emptyset$.*

Proof This follows from Note (3) at the beginning of this section. We must show that if aH and bH have at least one element in common then $aH = bH$. So suppose that there is an element c in $aH \cap bH$. By that note we have $aH = cH$ and $cH = bH$, as required.

There is an alternative way to present this proof, using the notion of equivalence relation. Define a relation R on G by xRy if and only if $y^{-1}x$ is in H. Then R is an equivalence relation:

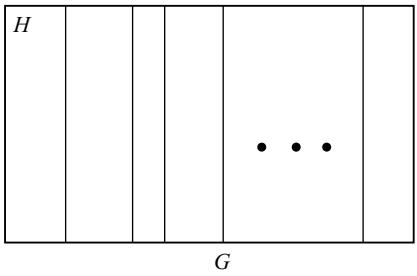


Fig. 5.2

clearly it is reflexive since e is in H;
it is symmetric since if $y^{-1}x$ is in H then so is $(y^{-1}x)^{-1} = x^{-1}y$;
it is transitive, since if $y^{-1}x$ and $z^{-1}y$ are in H then so is $z^{-1}x = (z^{-1}y)(y^{-1}x)$.

The equivalence class, $[x]$, of x is given by

$$\begin{aligned}[x] &= \{y \in G : yRx\} \\ &= \{y \in G : x^{-1}y \in H\} \\ &= \{y \in G : x^{-1}y = h \text{ for some } h \text{ in } H\} \\ &= \{y \in G : y = xh \text{ for some } h \text{ in } H\} \\ &= xH,\end{aligned}$$

so the result follows by Theorem 2.3.1. □

It follows that if H is a subgroup of the group G then every element belongs to one and only one left coset of H in G. Thus the (left) cosets of H in G form a partition of G. For instance in Example 4 above the cosets of the xy-plane in 3-space partition 3-space into a 'stack' of (infinitely many) parallel planes (note that two such planes are equal or have no point in common). In Example 1 we partitioned $\mathbb{Z}$ into the congruence classes of integers modulo n.

So we have the picture of G split up into cosets of the subgroup H (Fig. 5.2).

The next important point is that these pieces all have the same size.

Theorem 5.2.2 *Let H be a subgroup of the group G. Then each coset of H in G has the same number of elements as H.*

Proof (See Fig. 5.3.) Let aH be any coset of H in G. We show that there is a bijection between H and aH. Define the function $f : H \to aH$ by $f(h) = ah$.

(1) f is injective: for if $f(h) = f(k)$ we have $ah = ak$ and, multiplying on the left by a^{-1}, we obtain $h = k$ as desired.

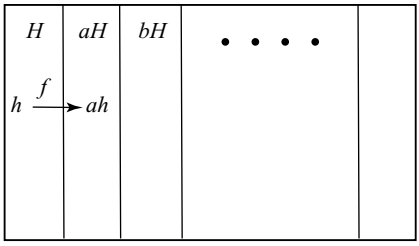


Fig. 5.3

(2) f is surjective: for if b is in aH then, by definition, b may be expressed in the form ah for some h in H; thus $b = f(h)$ is in the image of f, as required.

Therefore there is a bijection between H and aH, so they have the same number of elements (cf. Section 2.2), as claimed. (Observe that we did not assume that H is finite: the idea of infinite sets having the same numbers of elements was discussed in Section 2.2 where bijections were introduced and discussed.) □

Example In the example on p. 167, we showed, by a special case of the above argument, that the two cosets of the alternating group $A(n)$ in the symmetric group $S(n)$ have the same size.

This leads us to a key result, which is usually named after Joseph Louis Lagrange (1736–1813) who essentially established it. At least, he proved it in a special case: the general idea of a group did not emerge until around the middle of the nineteenth century. A second special case was shown by Cauchy. The general result was given by Jordan (who attributed it to Lagrange and Cauchy: the proof is the same in each case).

Definition Let G be a finite group. The **order** of G, $o(G)$, is the number of elements in G.

By Theorem 5.1.7, if G is cyclic, say $G = \langle g \rangle$, then the order of G is the order of the element g in the sense of the definition before 5.1.4 (the two uses of the term 'order' are distinct but related).

Theorem 5.2.3 (Lagrange's Theorem) *Let G be a finite group and let H be a subgroup of G. Suppose that H has m distinct left cosets in G. Then $o(G) = o(H) \cdot m$. In particular, the order of H divides the order of G.*

Proof We have only to put together the pieces that we have assembled. By Theorem 5.2.1, and since G is finite, G may be written as a disjoint union of

cosets of H: say

$$G = a_1H \cup a_2H \cup \ldots \cup a_mH$$

where $a_1 = e$ and $a_iH \cap a_jH = \emptyset$ whenever $i \neq j$.

By Theorem 5.2.2, the number of elements in each coset a_iH is $o(H)$ (that is, the number of elements in H). So, since the union is disjoint (by Theorem 5.2.1), the number of elements in G is $m \cdot o(H)$, as claimed. In particular, the number of elements of G is a multiple of the number of elements of H: in other words $o(H)$ must divide $o(G)$, as required. □

Comment Repeating the above proof using the distinct right cosets of H in G would also show that the number of these, n say, satisfies the same equation: $o(G) = n \cdot o(H)$. It therefore follows that $n = m$, so there are the same number of distinct right or left cosets of any given subgroup of a group. As we have seen, in general the list of distinct right cosets is not equal to the list of distinct left cosets, though the lists contain the same number of cosets.

We may quickly derive several important corollaries from Lagrange's Theorem.

Corollary 5.2.4 *Let g be an element of the group G. Then the order of g divides the order of the group G.*

Proof The order of g is equal to the number of elements in the cyclic subgroup which it generates (by Theorem 5.1.7). By Lagrange's Theorem, this number divides the number of elements in G. □

Corollary 5.2.5 *Let G be a group of prime order p. Then G is cyclic.*

Proof Let x be any element of G other than the identity. By Theorem 5.1.7, $\langle x \rangle$ is a subgroup of G and it certainly contains more then one element ($x \neq e$). By Lagrange's Theorem the number of elements in $\langle x \rangle$ divides p so, since it is greater than 1, it must be p. Thus $\langle x \rangle$ must be G. □

The next two corollaries of the result have been seen already in Section 1.6.

Corollary 5.2.6 (Fermat) *Let p be a prime number and a be any integer not divisible by p. Then $a^{p-1} \equiv 1$ mod p.*

Proof The group G_p of invertible congruence classes modulo p has $p - 1$ elements and consists of the congruence classes of integers not divisible by p.

Since the congruence class of a is in G_p it follows, by Corollary 5.2.4, that the order of a divides $p-1$. The result now follows by Theorem 5.1.4 since $[1]_p$ is the identity element of G_p. □

Corollary 5.2.7 (Euler) *Let n be any integer greater than 1 and let a be relatively prime to n. Then $a^{\phi(n)} \equiv 1$ mod n.*

Proof The proof is similar to that of Corollary 5.2.6 since the number of elements in G_n is $\phi(n)$. □

Remark Even from these few corollaries, one may appreciate that Lagrange's Theorem is very powerful: also illustrated is the strong connection between group theory and arithmetic.

Suppose that G is a group with n elements and let d be a divisor of n: it need not be the case that G has an *element* of order d (for instance, take G to be non-cyclic and take $d = n$). Indeed, the converse of Lagrange's Theorem is false in general. That is, if G is a group with n elements and if d is a divisor of n, G need not even have a *subgroup* with d elements. The simplest example here is the alternating group $A(4)$ of order 12 which has no subgroup with six elements. This is given as an exercise, with hints, in Section 5.3.

Corollary 5.2.4 shows an unexpected relationship between group theory and number theory. The order of a group influences its structure. This theme recurs throughout finite group theory which is, in fact, a very arithmetical subject. As an example of a way in which this relationship appears, we note that groups whose order is a power of a prime number p (p-groups) play a special role in the theory. The arithmetic also helps us understand the subgroup structure of a group. Although the converse of Lagrange's Theorem is false, each group of finite order divisible by a prime p does have certain important subgroups that are p-groups, and their existence (Sylow's Theorems) is one of the most significant results in the theory.

Exercises 5.2

1. Let G be the group G_{14} of invertible congruence classes modulo 14. Write down the distinct left cosets of the subgroup $\{[1]_{14}, [13]_{14}\}$.
2. Show that for $n \geq 3$, $\phi(n)$ is divisible by 2.
   [Hint: note that $(-1)^2 = 1$.]
3. Let G be the group $D(4)$ of symmetries of a square and τ be any reflection in G. Describe the left cosets of of the subgroup $\{1, \tau\}$ of G.

4. Let H be a subgroup of the group G and let a be an element of G. Fix an element b in aH (so b is of the form ah for some h in H). Show that
$$H = \{b^{-1}c : c \in aH\}.$$
5. Let G be the group G_{20}. What, according to Corollary 5.2.4, are the *possible* orders of elements of G and which of these integers are *actually* orders of elements of G?

5.3 Groups of small order

In this section we introduce the ideas of isomorphism and direct product. These will then be used to describe, up to isomorphism, all groups of order no more than 8.

Informally, we regard two groups G and H as being **isomorphic** ('of the same shape') if they can be given the 'same' multiplication table. More precisely, we require the existence of a bijection θ from G to H such that, if G is listed as $\{g_1, g_2, \ldots\}$ and if H is listed as $\{\theta(g_1), \theta(g_2), \ldots\}$ and if the multiplication tables are drawn up, then, if the (g_i, g_j) entry in the table for G is g_k, the $(\theta(g_i), \theta(g_j))$ entry in the table for H will be $\theta(g_k)$. This condition on the tables is simply that $\theta(g_i g_j) = \theta(g_i)\theta(g_j)$ for all i, j. (Notice that we used another Greek letter θ known as 'theta' to denote our bijection.)

Another way to understand the idea of two groups being isomorphic is to imagine the tables for the two groups G and H each being written on transparent slides. Do this using a different coloured pen for each element of the group G, replacing each occurrence of a given element g in the table (including the header row and column) by an appropriately coloured square. Use the colour associated with g for the element $\theta(g)$ in H and draw up another multi-coloured slide, replacing $\theta(g)$ wherever it occurs with a square coloured in the colour of g. Also, if necessary, rearrange the order of the entries in the header row and column so that they occur in the same order as on the slide with the table for G. The groups G and H are then isomorphic if the multi-coloured slide for G can be superimposed on the multi-coloured slide for H with no detectable differences. Thus any features of a group which may be determined from its multiplication table (such as being Abelian) must be shared with any isomorphic group.

There is actually no need to refer explicitly to the multiplication tables, and we make the following precise definition.

Definition Let G and H be groups. A function $\theta : G \longrightarrow H$ is an **isomorphism** (from G to H) if it is a bijection and if

for all $x, y \in G$ we have $\theta(xy) = \theta(x)\theta(y)$ $\quad (*)$

('θ preserves the group structure').

Groups G and H are **isomorphic** if there exists an isomorphism from G to H.

Notes (1) If G is any group then the identity function from G to itself is an isomorphism from G to G. That is, G is isomorphic to itself.

(2) If $\theta : G \longrightarrow H$ is an isomorphism then, as you are invited to verify, the inverse map $\theta^{-1} : H \longrightarrow G$ is an isomorphism. Thus, if G is isomorphic to H then so is H isomorphic to G.

(3) If $\theta : G \longrightarrow H$ and $\psi : H \longrightarrow K$ are isomorphisms, then the composition $\psi\theta : G \longrightarrow K$ is an isomorphism. Thus the relation of being isomorphic is transitive (ψ is the Greek letter 'psi').

(4) Taken together, (1), (2) and (3) show that the relation of being isomorphic is an equivalence relation.

Theorem 5.3.1 *Let θ be an isomorphism from G to H. Then $\theta(e_G)$ is the identity element of H and the inverse of $\theta(g)$ is $\theta(g^{-1})$. That is, $\theta(e_G) = e_H$ and $\theta(g^{-1}) = (\theta(g))^{-1}$.*

Proof By e_G we mean the identity element of G; similarly by e_H is meant the identity element of H. For every g in G, we have that

$$e_H \cdot \theta(g) = \theta(g) = \theta(e_G \cdot g) = \theta(e_G)\theta(g).$$

Using Theorem 5.1.1, it follows that $\theta(e_G)$ is the identity element e_H of H. Similarly, Theorem 5.1.1 applied to the equation

$$\theta(g) \cdot \theta(g)^{-1} = e_H = \theta(e_G) = \theta(gg^{-1}) = \theta(g)\theta(g^{-1})$$

shows that $\theta(g^{-1})$ is $\theta(g)^{-1}$. $\square$

Example 1 We have seen several examples of groups with four elements, namely $(\mathbb{Z}_4, +)$; the multiplicative group, G_5, of invertible congruence classes modulo 5; the subgroup of $S(4)$ consisting of the four permutations id, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3); the group of symmetries of a rectangle.

Of these, two are cyclic: G_5 is generated by $[2]_5$ and the additive group $\mathbb{Z}_4$ is generated by $[1]_4$. We may define a function θ from G_5 to $\mathbb{Z}_4$ by sending $[2]_5$ to $[1]_4$. If this function is to be structure preserving then it must be that $[4]_5 = [2]_5^2$ is sent to $[2]_4 = [1]_4 + [1]_4$ and that, more generally, that nth power of $[2]_5$ is sent to the nth power (or rather, sum, since $(\mathbb{Z}_4, +)$ is an additive group) of $[1]_4$. Thus, specifying where the generator is sent also determines where all its

powers are to be sent, assuming that the function is to satisfy condition $(*)$. Since the generators $[2]_5$ and $[1]_4$ have the same order, this function is well defined and, one may check, is an isomorphism (to see this directly, draw up the two group tables, then rearrange one of them using θ).

This shows that the groups G_5 and $\mathbb{Z}_4$ are isomorphic. The argument applies more generally to show that any two cyclic groups with the same number of elements are isomorphic (send a generator of one group to a generator of the other and argue as above).

For this reason, a cyclic group with n elements is often denoted simply by C_n (the fact that it is cyclic and has n elements determines it up to isomorphism). This notation is used when the group operation is written multiplicatively, whereas $\mathbb{Z}_n$ tends to be used in conjuction with additive notation.

The other two groups with four elements are isomorphic to each other, as may be seen be inspecting their multiplication tables (in Section 4.3). Specifically, an isomorphism from the given subgroup of $S(4)$ to the group of symmetries of a rectangle may be given by taking (1 2)(3 4) to σ, (1 3)(2 4) to R and (1 4)(2 3) to τ (and of course, id to e). However, neither of these is isomorphic to (either of) the cyclic groups. One way to see this is to observe that, in the second two examples, each element is its own inverse, but in the cyclic groups there are elements which are not their own inverse. (Any property defined solely in terms of the group operation must be preserved by any isomorphism.)

Example 2 We have seen two examples of non-Abelian groups with six elements: the symmetric group $S(3)$ and the dihedral group $D(3)$ of symmetries of an equilateral triangle. These are also isomorphic (see Example 1 in Section 4.3.4). It should be remarked that it may be very difficult to determine whether or not two (large) groups are isomorphic: even it they are isomorphic, there may be no 'obvious' isomorphism.

Example 3 You may check that the function $f : (\mathbb{R}, +) \longrightarrow (\mathbb{C}, \cdot)$ which takes $r \in \mathbb{R}$ to $e^{2\pi i r}$ satisfies the condition $(*)$ of the definition of isomorphism, but is not 1-1, since for any integer n one has $e^{2\pi i n} = 1$. Hence it is not an isomorphism.

Let us define the relation R on $\mathbb{R}$ by rRs if and only if $e^{2\pi i r} = e^{2\pi i s}$. Then R is an equivalence relation. It is straightforward to check that the equivalence class of 0 is a subgroup of $\mathbb{R}$: indeed it is just the set of all integers. The equivalence classes are just the cosets of this subgroup in $\mathbb{R}$. A group operation may be defined on the set G of equivalence classes (cosets), by setting $[r] + [s] = [r + s]$, where $[r]$ denotes the coset of r: you should check that this is well defined. Then define the function $g : G \longrightarrow \mathbb{C}$ by setting $g([r]) = f(r)$: again,

you should check that the definition of g does not depend on the representative chosen. Finally, let S be the image of the function g: it is the set of all complex numbers of the form $e^{2\pi ir}$, the circle in the complex plane with centre the origin and radius 1. Then g is an isomorphism from G to S.

Now we describe a way of obtaining new groups from old.

Definition Given groups G and H, the **direct product** $G \times H$ is the set of all ordered pairs (g, h) with g in G and h in H, equipped with the following multiplication

$$(g_1, h_1)(g_2, h_2) = (g_1h_1, g_2h_2).$$

Comment We have used the notation $X \times Y$ in Chapter 2 to denote the Cartesian product of X and Y. Here, our sets are actually groups so each has a operation which, although often written multiplicatively, should not be confused with the operation, $\times$, on sets. The direct product means 'the set of ordered pairs' and does not involve combining, in any way, the elements of G with those of H. Indeed, our ordered pair notation keeps elements of G apart from elements of H.

Theorem 5.3.2 *For any groups G and H, the direct product $G \times H$ is a group. In the case that G and H are finite, the order of this group is the product of the orders of G and H.*

Proof One checks the group axioms for $G \times H$: closure is clear; associativity follows from that for G and H; the identity is (e_G, e_H) and the inverse of (g, h) is (g^{-1}, h^{-1}). For the last part, see Exercise 2.1.8. □

Notes (1) Given groups G and H, the product groups $G \times H$ and $H \times G$ are isomorphic (define the isomorphism to take (g, h) to (h, g)).

(2) Forming direct products of groups is an associative operation in the sense that, given groups G, H, and K, there is an isomorphism from $(G \times H) \times K$ to $G \times (H \times K)$ (given by sending $((g, h), k)$ to $(g, (h, k))$) so we may write, without real ambiguity, $G \times H \times K$. The notations G^2, G^3, etc. are often used for $G \times G$, $G \times G \times G$, and so on.

Example 1 Let G and H both be the group G_3 of invertible congruence classes modulo 3. The group $G \times H$ has four elements:

$$([1]_3, [1]_3), ([1]_3, [2]_3), ([2]_3, [1]_3) \text{ and } ([2]_3, [2]_3).$$

It should be clear that $([1]_3, [1]_3)$ is the identity and that, for all a and b,

$$([a]_3, [b]_3)^2 = ([a^2]_3, [b^2]_3) = ([1]_3, [1]_3).$$

It is easy to check that $G \times H$ is Abelian.

Example 2 The direct product $S(3) \times S(3)$ has $36 (= 6 \times 6)$ elements. This group is not Abelian (since $S(3)$ is not Abelian).

Example 3 We may now explain a point which arose in Section 1.4. Let a be a generator for C_4 and let b be a generator for C_2. Then $C_4 \times C_2$ has eight elements:

$$(e, e), (e, b), (a, e), (a, b), (a^2, e), (a^2, b), (a^3, e) \text{ and } (a^3, b).$$

(Note that the 'e' appearing in the first coordinate is the identity element of C_4 whereas that appearing in the second coordinate is the identity element of C_2.)

It may be checked that this group is isomorphic to the group G_{20} by the function given by

$$(e, e) \to [1], (e, b) \to [11], (a, e) \to [3], (a, b) \to [13],$$
$$(a^2, e) \to [9], (a^2, b) \to [19], (a^3, e) \to [7] \text{ and } (a^3, b) \to [17].$$

This explains why the table for G_{20}, given in Section 1.4 splits into four blocks. The blocks are obtained by ignoring the first coordinate of the element of $C_4 \times C_2$ corresponding to a given element of G_{20}: so if we look only at the second coordinate then we obtain the structure of the multiplication table for C_2.

	(e,e)	(a,e)	(a^3,e)	(a^2,e)	(e,b)	(a,b)	(a^3,b)	(a^2,b)
(e,e)	(e,e)	(a,e)	(a^3,e)	(a^2,e)	(e,b)	(a,b)	(a^3,b)	(a^2,b)
(a,e)	(a,e)	(a^2,e)	(e,e)	(a^3,e)	(a,b)	(a^2,b)	(e,b)	(a^3,b)
(a^3,e)	(a^3,e)	(e,e)	(a^2,e)	(a,e)	(a^3,b)	(e,b)	(a^2,b)	(a,b)
(a^2,e)	(a^2,e)	(a^3,e)	(a,e)	(e,e)	(a^2,b)	(a^3,b)	(a,b)	(e,b)
(e,b)	(e,b)	(a,b)	(a^3,b)	(a^2,b)	(e,e)	(a,e)	(a^3,e)	(a^2,e)
(a,b)	(a,b)	(a^2,b)	(e,b)	(a^3,b)	(a,e)	(a^2,e)	(e,e)	(a^3,e)
(a^3,b)	(a^3,b)	(e,b)	(a^2,b)	(a,b)	(a^3,e)	(e,e)	(a^2,e)	(a,e)
(a^2,b)	(a^2,b)	(a^3,b)	(a,b)	(e,b)	(a^2,e)	(a^3,e)	(a,e)	(e,e)

Indeed, if we look at the blocks, then we have the multiplication table for the two cosets of the subgroup $C_4 \times \{e\}$ in $C_4 \times C_2$ (cf. Example 3 on p. 221).

Example 4 When both G and H are Abelian, we often write the group operation in $G \times H$ as addition. For example, the group $\mathbb{Z}_2 \times \mathbb{Z}_2$ has four elements

$$([0]_2, [0]_2), ([0]_2, [1]_2), ([1]_2, [0]_2) \text{ and } ([1]_2, [1]_2).$$

Since

$$\begin{aligned}([a]_2, [b]_2) + ([c]_2, [d]_2) &= ([a + c]_2, [b + d]_2) \\ &= ([c + a]_2, [d + b]_2) = ([c]_2, [d]_2) + ([a]_2, [b]_2),\end{aligned}$$

we see that $\mathbb{Z}_2 \times \mathbb{Z}_2$ is an Abelian group. In fact, $\mathbb{Z}_2 \times \mathbb{Z}_2$ is isomorphic to the group in Example 1 as well as to the group of symmetries of a rectangle. This group with four elements is often referred to as the **Klein four group** and is denoted $\mathbb{Z}_2 \times \mathbb{Z}_2$ or $C_2 \times C_2$ depending on whether we wish to use additive or multiplicative notation.

Theorem 5.3.3 *Let m and n be relatively prime integers. Then the direct product $C_m \times C_n$ is cyclic.*

Proof Let a and b be generators for C_m and C_n respectively. So, by 5.1.7, the order of a is m and that of b is n. Then, for any integer k,

$$(a, b)^k = (a^k, b^k)$$

(this is proved by induction, using the definition of the group operation in the direct product). Thus, if $(a, b)^k = (e, e) = (a, b)^0$ then we have, by Theorem 5.1.4, that both m and n divide k. Since m and n are relatively prime, mn divides k by Theorem 1.1.6.

We also have $(a, b)^{mn} = (a^{mn}, b^{mn}) = ((a^m)^n, (b^n)^m) = (e, e)$. It follows that the order of (a, b) is mn. Thus the distinct powers of (a, b) exhaust the group $C_m \times C_n$ (which has mn elements) and so the group is indeed cyclic. □

We now start our classification of groups of small orders.

Groups of order 1 Any group contains an identity element e, so if G has only one element then G consists of only the identity element. Clearly any two such groups are isomorphic!

Groups of order 2 If G has two elements, we must have $G = \{e, g\}$ for some g different from e. Since G is a group and so is closed under the operation, g^2 is in G. Now, g^2 cannot be g, for $g^2 = g$ implies (multiply each side by g^{-1}) $g = e$. It must therefore be that $g^2 = e$. This lets us construct the group

table and also shows that there is only one possibility for the shape of this table. Hence there is (up to isomorphism) just the one group of order 2.

	e	g
e	e	g
g	g	e

Groups of order 3 Suppose that G has three different elements e (identity), g and h. We must have $gh = e$ (otherwise $gh = g$ or $gh = h$ and cancelling gives a contradiction). Similarly $hg = e$. This is enough to allow us to construct the group table. For example, g^2 is not e (otherwise $g^2 = e = gh$ so cancelling gives $g = h$, contrary to assumption) nor is it g (otherwise $g = e$) and so g^2 must be h. Thus G has the table shown.

	e	g	g^2
e	e	g	g^2
g	g	g^2	e
g^2	g^2	e	g

Remark Since 2 and 3 are prime numbers, the above two cases are, in fact, covered by the Corollary 5.2.5 to Lagrange's Theorem and the remarks on pp. 220 and 221.

Groups of order 4 First suppose that there is an element g in G of order 4. Then G must consist of $\{e, g, g^2, g^3\}$, and the multiplication table is constructed easily (the first table below): we note that G is cyclic.

If there is no element of order 4 then, by Corollary 5.2.4, each non-identity element of $G = \{e, g, h, k\}$ must have order 2. Also, by the kind of cancelling argument that we have already used, it must be that $gh = k$. The following result is of general use.

Theorem 5.3.4 *Let G be a group in which the square of every element is* 1. *Then G is Abelian.*

Proof For all g in G, we have that $g^2 = e = gg$. Thus every element is its own inverse. Since the inverse of xy is also $(xy)^{-1} = y^{-1}x^{-1}$ by 5.1.2, we have

$$xy = (xy)^{-1} = y^{-1}x^{-1} = yx$$

and so G is Abelian. □

This applies to our non-cyclic group $G = \{e, g, h, k\}$ and we construct the group table (the second below) easily using the facts that rows and columns of group tables can have no repeated elements. We have shown that any group with four elements is isomorphic to one of the two groups given by the tables below.

	e	g	g^2	g^3
e	e	g	g^2	g^3
g	g	g^2	g^3	e
g^2	g^2	g^3	e	g
g^3	g^3	e	g	g^2

	e	g	h	k
e	e	g	h	k
g	g	e	k	h
h	h	k	e	g
k	k	h	g	e

In other terms, we have shown that a group of order 4 is either cyclic or isomorphic to the Klein four group $C_2 \times C_2$.

Groups of order 5 Since 5 is a prime, we deduce (by Corollary 5.2.5) that G is cyclic, isomorphic to C_5, and consists of the powers of an element of order 5. Hence we can draw up the group table.

	e	g	g^2	g^3	g^4
e	e	g	g^2	g^3	g^4
g	g	g^2	g^3	g^4	e
g^2	g^2	g^3	g^4	e	g
g^3	g^3	g^4	e	g	g^2
g^4	g^4	e	g	g^2	g^3

Groups of order 6 If G contains an element of order 6 then there is no room in G for anything other than the powers of this element and so G is cyclic, isomorphic to C_6.

By Lagrange's Theorem, the only possible orders of elements of G are 1, 2, 3 and 6. Suppose then that G does not contain an element of order 6. If G contained no element of order 3 then all the non-identity elements of G would have to have order 2 and then, by Theorem 5.3.4, G would be Abelian. If that were the case, let a and b be non-identity elements of G. Then ab cannot be e, a or b (by 'cancelling' arguments). It follows that $\{e, a, b, ab\}$ is a subgroup of G (for this set is closed under products and inverses). But 4 does not divide 6, so Lagrange's Theorem (5.2.3) says that this is impossible.

Therefore G does have an element a (say) of order 3. Thus we have three of the elements of G: e, a, a^2. Let b be any other element of G. Then the six elements e, a, a^2, b, ba, ba^2 must be distinct: just note that any equation between them, on cancelling, leads to something contrary to what we have

assumed. For example, if we have already argued that e, a, a^2, b and ba are distinct, then ba^2 is different from each:

if $ba^2 = e$, then b would be the inverse of a^2, so $b = a$;
if $ba^2 = a$, then $ba = e$ and b would be $a^{-1} = a^2$;
if $ba^2 = a^2$, then $b = e$;
if $ba^2 = b$, then $a^2 = e$; and
if $ba^2 = ba$, then $a = e$.

In a similar way, we can show that b^2 must be e, since if b^2 were equal to b, ba or ba^2, we could deduce that the elements would not be distinct. If b^2 were a or a^2, it would follow that the powers of b $(b, b^2, \ldots, b^5)$ would be distinct and hence G would be cyclic. Similar arguments show that ab must be ba^2 and that in fact the multiplication table of G is that shown on the right below.

	e	g	g^2	g^3	g^4	g^5
e	e	g	g^2	g^3	g^4	g^5
g	g	g^2	g^3	g^4	g^5	e
g^2	g^2	g^3	g^4	g^5	e	g
g^3	g^3	g^4	g^5	e	g	g^2
g^4	g^4	g^5	e	g	g^2	g^3
g^5	g^5	e	g	g^2	g^3	g^4

	e	a	a^2	b	ba	ba^2
e	e	a	a^2	b	ba	ba^2
a	a	a^2	e	ba^2	b	ba
a^2	a^2	e	a	ba	ba^2	b
b	b	ba	ba^2	e	a	a^2
ba	ba	ba^2	b	a^2	e	a
ba^2	ba^2	b	ba	a	a^2	e

The group on the left is cyclic of order 6, while that on the right is (necessarily) isomorphic to the symmetric group $S(3)$. An isomorphism f from the group on the right to $S(3)$ is given by

$$f(e) = \text{id}; \qquad f(a) = (1\ 2\ 3); \qquad f(a^2) = (1\ 3\ 2);$$
$$f(b) = (1\ 2); \qquad f(ba) = (2\ 3); \qquad f(ba^2) = (1\ 3).$$

Therefore a group of order 6 is isomorphic either to the cyclic group of order 6 or to the group $S(3)$.

Groups of order 7 As in the case of orders 3 and 5 we see that the only possibility is a cyclic group, C_7, the cyclic group with 7 elements, since 7 is prime. Drawing up the group table is left as an exercise.

Groups of order 8 At this point, we will just present the answer and leave the details of the calculations to Exercise 5.3.10 at the end of the section. By the kind of analysis we used for groups of order 6, it can be shown that there are five different types of group with eight elements. Three of these are Abelian

and are C_8, $C_4 \times C_2$ and $C_2 \times C_2 \times C_2$. There are two types of non-Abelian group, the dihedral group $D(4)$ and the quaternion group $\mathbb{H}_0$. The tables for the last two can be found in the solution for Exercise 4.3.7, and in Example 5 of Section 4.3.1, respectively.

In the case of groups of order 8, it can be seen that the techniques we have developed become rather stretched and other methods are required to make progress on the problem of finding all groups with a given number of elements. The interested reader should consult one of the abundance of more advanced books devoted to group theory to see what techniques are available to study groups in general. However, it may be of interest to say a little more about the classification problem for finite groups.

In some sense, every finite group is built up from 'simple groups'. In order to define this term, we need to introduce the notion of a normal subgroup of a group. A subgroup N of a group G is **normal** if, for each g in G the left coset gN is equal to the right coset Ng. A group G is **simple** if the only normal subgroups of G are G itself and the trivial subgroup $\{e\}$. The importance of normal subgroups is on account of the following (cf. Example 3 before Theorem 5.3.2 and Example 3 before Theorem 5.3.3).

If N is a normal subgroup of the group G then the set of cosets of N in G may be turned into a group by defining $(gN) \cdot (hN) = ghN$ (normality of N is needed for this multiplication to be well defined). This group of cosets is denoted by G/N: it is obtained from G by 'collapsing' the normal subgroup to a single element (the identity) of the group G/N. One may say that the group G is built up from the normal subgroup N and the group G/N. So, if a group is not simple, then it may in some sense be decomposed into two smaller (so simpler) pieces. But a simple group cannot be so decomposed. Therefore the simple groups are regarded as the 'building blocks' of finite groups, in a way analogous to that in which the prime numbers are the 'building blocks' of positive integers.

The complete list of finite simple groups is known and its determination, which was essentially completed in the early 1980s, was one of the great achievements in mathematics. There are two aspects to this result. One is the production of a list of finite simple groups and the other is the verification that every finite simple group is on the list.

Most of the finite simple groups fall into certain natural infinite families of closely related groups such as groups of permutations or matrices: for instance, the alternating $A(n)$ for $n \geq 5$ are simple. But five anomalous, or sporadic, simple groups, groups which do not fit into any infinite family, were discovered by Mathieu between 1860 and 1873. No more sporadic simple groups were

found for almost a hundred years, until Janko discovered one more in 1966. Between then and 1983 a further 20 were found, bringing the total of sporadic simple groups to 26. The last found and largest of these is the so-called Monster (or Friendly Giant). The possible existence of this simple group was predicted in the mid-1970s and a construction for it was given by Robert Greiss in 1983. This is a group in which the number of elements is

$$2^{46} \times 3^{20} \times 5^9 \times 7^6 \times 11^2 \times 13^3 \times 17 \times 19 \times 23 \times 29 \times 31 \times 41 \times 47 \times 59 \times 71.$$

Clearly, one cannot draw up the multiplication table for such a group, and it is an indication of the power of techniques in current group theory that a great deal is understood about this largest sporadic simple group.

With Greiss' construction in 1983, the last finite simple group has been found, and thus the classification is complete, since all other possibilities for new finite simple groups have been excluded. It is estimated that over 200 mathematicians have contributed to this classification of the finite simple groups, and that the detailed reasoning to support the classification occupies over 15 000 printed pages. Since the first proof a number of mathematicians have been working to simplify the details.

Exercises 5.3

1. For each of the following groups with four elements, determine whether it is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$ or $\mathbb{Z}_4$:
   (i) the multiplicative group G_8 of invertible congruence classes modulo 8;
   (ii) the cyclic subgroup $\langle \rho \rangle$ of $D(4)$ generated by the rotation ρ of the square through $2\pi/4$;
   (iii) the groups with multiplication tables as shown (where the identity element does not necessarily head the first row and first column).

	a	b	c	d
a	d	c	a	b
b	c	d	b	a
c	a	b	c	d
d	b	a	d	c

	a	b	c	d
a	b	a	d	c
b	a	b	c	d
c	d	c	b	a
a	c	d	a	b

2. Consider the subgroup of $S(4)$ and the group of symmetries of the rectangle discussed after Theorem 5.3.1. There we defined one isomorphism between these groups but this is not the only one. Find all the rest.
   [Hint: choose any two elements of order 2; what are the restrictions on where an isomorphism can take them? Having determined where these two

elements are sent by the isomorphism, is there any choice for the destinations of the other two elements?]

3. Let G be any group and let g be an element of G. Define the function $f : G \longrightarrow G$ by $f(a) = g^{-1}ag$ $(a \in G)$ (thus f takes every element to its conjugate by g). Show that f is an isomorphism from G to itself. Show, by example, that f need not be the identity function.
4. Give an example of cyclic groups G and H such that $G \times H$ is not cyclic.
5. Show that $G \times H$ is Abelian if and only if both G and H are Abelian.
6. Show that the set $\{(g, e) : g \in G\}$ forms a subgroup of $G \times H$.
7. Write down the multiplication tables for the direct products
   (i) $\mathbb{Z}_4 \times \mathbb{Z}_2$;
   (ii) $G_5 \times G_3$;
   (iii) $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$;
   (iv) $G_{12} \times G_4$.
   Which of the above groups are isomorphic to each other?
8. Let G be a group with six elements and let H be a group with fourteen elements. What are the possible orders of elements in the direct product $G \times H$?
9. Use the classification of groups with six elements to show that $A(4)$ has no subgroup with 6 elements.
   [Hint: check that the product of any two elements of $A(4)$ of order 2 has order 2.]
10. Let G be a non-Abelian group with eight elements. Show that G has an element a say of order 4. Let b be an element of G which is not e, a, a^2 or a^3. By considering the possible values of b^2 and of ba and ab, show that G is isomorphic either to the dihedral group or to the quaternion group.

5.4 Error-detecting and error-correcting codes

Messages sent over electronic and other channels are subject to distortions of various sorts. For instance, messages sent over telephone lines may be distorted by other electromagnetic fluctuations; information stored on a disc may become corrupted by strong magnetic fields. The result is that the message received or read may be different from that originally sent or stored. Extreme examples are the pictures sent back from space probes, where a very high error rate occurs.

It is therefore important to know if an error has occurred in transmission: for then one may ask that the message be repeated. In some circumstances it may be impossible or undesirable for the message to be repeated. In that case, the message should carry a certain degree of redundancy, so that the original message may be reconstructed with a high degree of certainty. The way to do

this is to add a number of check symbols to the message so that errors may be detected or even corrected. In general, the greater the number of check symbols, the more unlikely it is that an error will go undetected.

Notice that the codes discussed here are not designed to prevent confidential information from being read (in contrast to the public key codes of Chapter 1). The object here is to ensure the accuracy of the message after transmission.

Another point which should perhaps be made explicit here is that there can be no method which reconstructs a distorted message with 100 per cent accuracy. When we say that a received message m is 'corrected' to m_1, we mean that it is *most likely* that m_1 was the message originally sent. If there is a low frequency of errors, then the probability that an error is wrongly 'corrected' may be made extremely small. This point is discussed further in Example 2 below.

We introduce the general idea of a coding function and discuss the concepts of error detection and correction. Then we specialise to linear codes. One way to produce codes of this sort is to use a generator matrix. This matrix tells us how to build in some redundancy by adding check digits to the original message. The subsequent correction of the message can be carried out using a coset decoding table. In producing this table, we again encounter the idea of a coset which was fundamental to the proof of Lagrange's Theorem.

Example A well known example of error correction is provided by the ISBN (International Standard Book Number) of published books. This is a sequence of nine digits $a_1a_2 \ldots a_9$ where each a_i is one of the numbers $0, 1, \ldots, 9$, together with a check digit which is one of the symbols $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ or X (with X representing 10). This last digit is included so as to give a check that the previous 9 digits have been correctly transcribed, and is calculated as follows. Form the integer $n = 10a_1 + 9a_2 + 8a_3 + \cdots + 2a_9$ and reduce this modulo 11 to obtain an integer b between 1 and 11 inclusive. The check digit is obtained by subtracting b from 11. Thus, for the number 0 521 35938, b is obtained by reducing

$$n = 0 + 45 + 16 + 7 + 18 + 25 + 36 + 9 + 16$$

modulo 11: the result is 7 and the check digit is $11 - 7 = 4$. The ISBN corresponding to 0 521 35938 is therefore 0 521 35938 4. If a librarian or bookseller made a single error in copying this number (say the IBSN was written as 0 521 35958 4) then the check digit obtained from the first nine digits would not be 4. (This is because an error of size k in position i would produce an error of k times $11 - i$ in n. Since 11 is prime, the error thereby introduced into the sum would not be divisible by 11.) This is an example of a code which *detects* a single error. Since there is no way of telling where the error is, the code is not error *correcting*.

The bar code used on many products also contains a check digit.

In the ISBN code, numbers are represented in the decimal system but we will consider from now on information which is stored or transmitted in binary form (of course the same general principles apply to other cases). This includes any information handled by a computer. Most other forms may be converted to binary: for example, English text may be converted by replacing each letter, numeral, space or punctuation mark by a suitable binary-based code (such as ASCII) for it. So from now on, we will consider only codes which apply to strings of 0s and 1s. We will think of the set $\{0,1\}$ as coming equipped with the operation of addition (and multiplication) mod 2: it is customary in this context to write $\mathbf{B}$ instead of $\mathbb{Z}_2$ (since $\mathbb{Z}_2$ with the operations of addition and multiplication is a Boolean ring, see Exercise 4.4.16).

Definition A **word** of **length** n is a string of n binary digits. Thus 0001, 1110 and 0000 are words of length 4. We shall think of words of length n as members of $\mathbf{B}^n$, the Cartesian product of n copies of the binary set $\mathbf{B}$ regarded as an Abelian group under addition. In this notation, if w is a word in $\mathbf{B}^n$ and x is in $\mathbf{B}$ we use wx to denote the word of length $n+1$ with its first n 'letters' being those of w and its last letter being x. This should not be confused with the product of w by x.

We formalise the idea of check symbols as follows. Suppose our original messages are composed of words of length m; we choose a 'coding function' $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ and, instead of sending a word w, we send the word $f(w)$. Thus the messages we send are composed of words of length n (rather than m). Any word of length n in the image of f is called a **codeword**.

There is an obvious constraint on a suitable coding function f: f should be injective, otherwise there would be two different words of length m that would be sent as the same word of length n. Note that this means that n should be greater than or equal to m and, in practice, strictly greater than m since we wish to add some check digits ($n - m$ of them).

Here are two examples of coding functions.

Example 1 Define $f : \mathbf{B}^m \longrightarrow \mathbf{B}^{m+1}$ by $f(w) = wx$ where x is 0 if the number of non-zero digits in w is even, and x is 1 if the number of non-zero digits in w is odd. To give a more specific example, take $m = 3$: there follow the eight words in $\mathbf{B}^3$ and beneath each is its image under f.

000	001	010	011	100	101	110	111
0000	0011	0101	0110	1001	1010	1100	1111

The last digit is a parity-check digit: any correctly transmitted word has an even number of 1s in it. Therefore this code enables one to detect any single error in the transmission of a codeword since, if a single digit is changed, the word received will then have an odd number of 1s in it and so not be a *codeword*. In fact any odd number of errors will be detected, but an even number of errors will fail to be detected. Another point about this code is that it does not allow one to correct an error without re-transmission of the word.

Example 2 Define $f : \mathbf{B}^m \longrightarrow \mathbf{B}^{3m}$ by $f(w) = www$; thus the word is simply repeated three times. So, for example, if $m = 6$ and if $w = 101111$ then $f(w) =$ 101111101111101111. You should convince yourself that this code will detect any single error or any two (non-cancelling) errors. For instance if $f(w)$ as above is received as 100111101011101111 (with two errors) then we can see at once that an error has occurred in transmission since the received message is not a six-letter word three times repeated. However this code does not necessarily detect three errors. If $f(w)$ were received as 001111001111001111 (three errors) then it looks as if the original word w was 001111, whereas the original word was 101111.

It should also be noted that although two errors can be *detected*, this code can *correct* only one error. For instance if $f(w)$ were received as 101011101011101111 then we could consider it most likely that the original word was 101011, not 101111. Let us be more explicit.

Suppose that www is the word sent and m is the word received: breaking it into three blocks of six letters each, we write $m = abc$ where a, b and c are words of length 6 and 'abc' means the word whose first six letters are those of a, whose next six letters are those of b and whose last six letters are those of c. If no errors have been made in transmission then $a = b = c = w$. If one error has been made, then two of a, b, and c are equal to each other (so, necessarily to w), so we correct the message and conclude (correctly) that the original word is w. If two errors have been made then it could happen that $abc = w'ww'$ (say): we would then conclude (incorrectly) that the original word was w'. We say, therefore, that this code can correct one error (but not two).

Thus we correct errors on the basis of 'most likely message to have been sent, on the assumption that errors occur randomly'. Suppose, for illustration, that the probability that a given digit is transmitted wrongly is 1 in 1000 (0.001). Then the probability that a single transmitted word (18 digits) contains a single error is 0.017 696 436 (that is, about 1 in 60); the probability that a given word contains two (respectively three or more) errors is 0.000 150 570 (0.000 000 806). It follows that the probability of incorrectly 'correcting' a word containing an error is less than one in ten thousand and the probability of failing to detect that a received word is erroneous is about one in one hundred million (note

that, even given that three errors have occurred, it is a small probability that the result consists of a six-letter word three times repeated). A book of 1000 pages, 40 lines to a page, around 60 characters (including spaces) to a line, contains about 2 400 000 characters. Six binary digits are easily sufficient to represent the alphabet plus numerals and punctuation, so the book may be represented by 2 400 000 binary words of length 6. So, with the above likelihood of error, the probability that even just one character of the book is transmitted wrongly and the error not detected is about 1 in 40.

The code in Example 2 above is superior to that in Example 1 in that it can detect up to two errors (rather than only one) and can even correct any single error. On the other hand it requires the sending of a message three times as long as the original one, whereas the first code involves only a slight increase in the length of message sent. Most of our examples below will be more efficient than this second one.

Definitions The **weight** of a binary word w, $\mathrm{wt}(w)$, is defined to be the number of 1s in its binary expression. Thus, for example

$$\mathrm{wt}(001101) = 3; \qquad \mathrm{wt}(000) = 0; \qquad \mathrm{wt}(111) = 3.$$

The **distance** between binary words v, w of the same length is defined to be the weight of their difference:

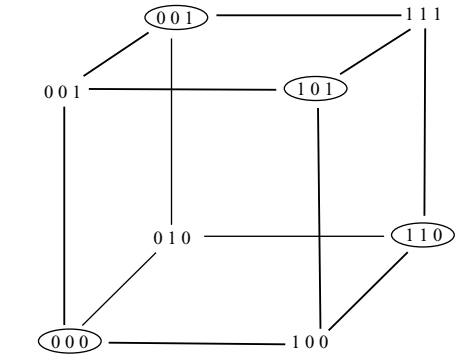
$$d(v, w) = \mathrm{wt}(v - w).$$

You should note that, since we are working in a product of copies of **B** (in which $+1 = -1$), every element is its own additive inverse and so, for example $v - w = v + w$. Hence $d(v, w) = \mathrm{wt}(v + w)$. It follows also that the ith entry of $v + w$ is 0 if the ith entries of v and w are the same, and is 1 if v and w differ at their ith places. Hence the distance between v and w is simply the number of places at which they differ.

Example

$$d(010101, 101000) = \mathrm{wt}(010101 + 101000) = \mathrm{wt}(111101) = 5;$$
$$d(1111, 0110) = \mathrm{wt}(1111 + 0110) = \mathrm{wt}(1001) = 2;$$
$$d(w, w) = \mathrm{wt}(w + w) = 0 \text{ for any word } w.$$

If w is a word that is transmitted and is received as v then the number of errors which have occurred in transmission is (ignoring errors which cancel) the distance between v and w (for the alteration of a single digit of a word results in a word at a distance of 1 from the original word). Therefore a good coding



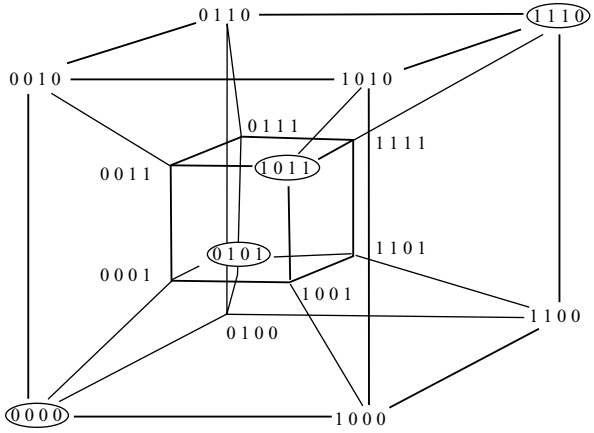


Fig. 5.4

function $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ will be one which maximises the minimum distance between different codewords.

We can illustrate this by drawing a graph, rather like those in Section 2.3 but with undirected edges. For the vertices of the graph, we take all the binary words of length n, and we join two vertices by an edge if the distance between them is 1 (that is, if an error in a single digit can convert one to the other). Then the distance between two words is the number of edges in a shortest path from one to the other. A good coding function is one for which the codewords are well 'spread' through this graph.

We show a couple of examples (Fig. 5.4), in which the codewords have been ringed. We have limited ourselves to words of length 3 and 4, since words of length n would most naturally be represented as the vertices of an

n-dimensional cube, and representing a five-dimensional cube on a piece of paper would be messy. The first example shows the codewords for the parity-check code $f : \mathbf{B}^2 \longrightarrow \mathbf{B}^3$. The second shows the codewords for a coding function $f : \mathbf{B}^2 \longrightarrow \mathbf{B}^4$.

Theorem 5.4.1 *Let $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ be a coding function. Then f allows the detection of k or fewer errors if and only if the minimum distance between distinct codewords is at least $k + 1$.*

Proof If a word w is obtained from a codeword by making k (or fewer) changes then w cannot be another codeword if the minimum distance between distinct codewords is $k + 1$. Thus the code will detect these errors. Conversely, if the code detects k errors, then no two codewords can be at a distance k from each other (for then k errors could convert one codeword to another and the change would not be detected). □

Theorem 5.4.2 *Let $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ be a coding function. Then f allows the correction of k or fewer errors if and only if the minimum distance between distinct codewords is at least $2k + 1$.*

Proof If the distance between the codewords v and w is $2k + 1$ then $k + 1$ errors in the transmission of v will indeed be detected. But the resulting word may be closer to w than to v, and so any attempt at error correction could result in the (incorrect) interpretation that w was more likely than v to have been the word sent. □

Example 3 Suppose we define the coding function $f : \mathbf{B}^4 \longrightarrow \mathbf{B}^9$ by setting $f(w) = www$ where x is 0 or 1 according as the weight of the word w is even or odd (so our coding function repeats the word and also has a parity-check digit). Opposite each word w in $\mathbf{B}^4$ we list $f(w)$:

0000	000000000	0001	000100011
0010	001000101	0011	001100110
0100	010001001	0101	010101010
0110	011001100	0111	011101111
1000	100010001	1001	100110010
1010	101010100	1011	101110111
1100	110011000	1101	110111011
1110	111011101	1111	111111110

You may check, by computing $d(u, v)$ for all $u \neq v$, that the minimum distance between codewords is 3 and so, by Theorems 5.4.1 and 5.4.2, the code detects up to two errors and can correct any single error.

Checking the minimum distance between codewords is a tedious task: but it can be circumvented for certain types of codes by using a little theory.

Definition Let $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ be a coding function. We say that this gives a **linear code** if the image of f forms a subgroup of $\mathbf{B}^n$. (These codes are also referred to as group codes, but the word 'linear' helps to remind us that the group operation is addition.)

What do we have to check in order to show that we have a linear code? Since the group operation is addition, we must show that if words u and v are in the image of f then so is the word $u + v$. We do not have to check that if u is a codeword then $-u$ also is a codeword since in $\mathbf{B}^n$ every element is self-inverse! ($-u = u$): also '$\mathbf{0}$' may be obtained as $u + u$ for any u in the image of f (where $\mathbf{0}$ denotes the codeword with all entries '0').

One advantage of linear codes is that the minimum distance between codewords is relatively easily found.

Theorem 5.4.3 *Let $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ be a linear code. Then the minimum distance between distinct codewords is the lowest weight of a non-zero codeword.*

Proof Let d be the minimum distance between distinct codewords: so $d(u, v) = d$ for some codewords u, v. Let x be the minimum weight of a non-zero codeword: so $\mathrm{wt}(w) = x$ for some codeword w. Then, since d is the minimum distance between codewords and w and $\mathbf{0}$ are codewords, we have:

$$d \leq d(w, \mathbf{0}) = \mathrm{wt}(w + \mathbf{0}) = \mathrm{wt}(w) = x.$$

On the other hand

$$d = d(u, v) = \mathrm{wt}(u + v).$$

Since we have a linear code, $u + v$ is a (non-zero) codeword so, by minimality of x, $\mathrm{wt}(u + v) \geq x$: thus $d \geq x$. Hence $d = x$ as claimed. □

We now present a method of producing linear codes.

Definition Let m, n be integers with $m < n$. A **generator matrix** G is a matrix with entries in $\mathbf{B}$ and with m rows and n columns, the first m columns of which

form the $m \times m$ identity matrix $\mathbf{I}_m$. We may write such a matrix as a partitioned matrix: $G = (\mathbf{I}_m A)$ where A is an $m \times (n - m)$ matrix.

For instance, the following are generator matrices (of various sizes):

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}; \quad \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix};$$

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}; \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Given a generator matrix G with m rows and n columns, we may define the corresponding coding function $f = f_G : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ by treating the elements of $\mathbf{B}^m$ as row vectors and setting $f_G(w) = wG$ for w in $\mathbf{B}^m$.

For example, suppose that G is the second matrix above: so $f_G : \mathbf{B}^3 \longrightarrow \mathbf{B}^4$. We have, for instance

$$f_G(011) = (011) \cdot G = (011) \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} = (0110).$$

Observe that if $f = f_G : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ arises from a generator matrix then the first m entries of the codeword $f(w)$ form the original word w: thus $f(w)$ has the form wv, where v contains the 'check digits'.

Theorem 5.4.4 *Let G be a generator matrix. Then f_G is a linear code and, in fact, $f_G(v + w) = f_G(v) + f_G(w)$ for all words v, w.*

Proof We have $f_G(\mathbf{0}_m) = \mathbf{0}_m G = \mathbf{0}_n$ where $\mathbf{0}_k$ denotes the k-tuple with all entries '0': so the set of codewords is non-empty. Suppose that u, u' are codewords: say

$$u = f_G(v) \text{ and } u' = f_G(w).$$

Then

$$\begin{aligned} f_G(v + w) &= (v + w)G \\ &= vG + wG \text{ (matrix multiplication distributes over addition)} \\ &= f_G(v) + f_G(w) \\ &= u + u'. \end{aligned}$$

Thus the set of codewords is closed under addition. We have already noted that we do not need to check for inverses because every element is its own inverse. Thus the set of codewords is a group, as required. □

Referring back to Example 3 above we see that we can use Theorem 5.4.3 to derive that the minimum distance between (distinct) codewords is 3, the minimum weight of a non-zero codeword. We can justify the use of Theorem 5.4.3 by checking that the code is given by the generator matrix below (then appealing to Theorem 5.4.4):

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Examples 1 and 2 on pp. 232 and 233 are also linear codes obtained from generator matrices. In Example 1, the matrix G is obtained from the $m \times m$ identity matrix by adding a column of 1s to the right of it. In Example 2, the generator matrix simply consists of three $m \times m$ identity matrices placed side by side. Let us consider some further examples. After giving the generator matrix G we list the words of $\mathbf{B}^m$ beside the corresponding codewords in $\mathbf{B}^n$.

Example 1 Consider the coding function $f : \mathbf{B}^2 \longrightarrow \mathbf{B}^4$ with generator matrix

$$G = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \qquad \begin{array}{ll} 00 & 0000 \\ 01 & 0101 \\ 10 & 1011 \\ 11 & 1110 \end{array}$$

The minimum distance between codewords is 2 (being the minimum weight of a non-zero codeword), so the code can detect one error but cannot correct errors.

Example 2 Next take $f : \mathbf{B}^2 \longrightarrow \mathbf{B}^5$ with

$$G = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix} \qquad \begin{array}{ll} 00 & 00000 \\ 01 & 01011 \\ 10 & 10110 \\ 11 & 11101 \end{array}$$

The minimum distance between codewords is 3, so the code can detect two errors and correct one error.

Example 3 Consider the coding function $f : \mathbf{B}^3 \longrightarrow \mathbf{B}^6$ with

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

000	000000
001	001111
010	010101
011	011010
100	100111
101	101000
110	110010
111	111101

The minimum distance between codewords is 2 (being the minimum weight of a non-zero codeword), so the code can detect one error but cannot correct even all single errors.

Example 4 Finally consider the coding function $f : \mathbf{B}^3 \longrightarrow \mathbf{B}^{10}$ given by

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

000	0000000000
001	0010111001
010	0101011010
011	0111100011
100	1001101100
101	1011010101
110	1100110110
111	1110001111

The minimum distance between codewords is 5, so the code can detect four errors and correct two errors. Note how much better this is than the code which simply repeats the word a total of four times: the latter code uses more digits (12 instead of 10) yet detects and corrects fewer errors (it detects three and can correct one error).

If you are familiar with 'linear algebra', you may have realised that in our use of matrices we are essentially doing linear algebra over the field $\mathbf{B} = \mathbb{Z}_2$ of 2 elements. More precisely, in the terminology of Section 4.4, we are regarding $\mathbf{B}^n$ as a vector space over the field $\mathbf{B} = \mathbb{Z}_2$.

Now we describe how to arrange the work of detecting and correcting errors in any received message. *Suppose throughout that we are using a linear code* $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$. Let W be the set of codewords in $\mathbf{B}^n$ (a subgroup of $\mathbf{B}^n$).

Suppose that the word w is sent but that an error occurs in, say, the last digit, resulting in the word v being received. So v agrees with w on each digit but

the last, where it differs from w: that is $v = e_n + w$ where e_n is the word of length n which has all digits '0' except the last, which is '1'. Thus the set of words which may be received as the result of a single error in the last digit is precisely the set $e_n + W$. In the language of group theory, this is precisely the coset of e_n with respect to the subgroup W of $\mathbf{B}^n$. (Since the group operation is addition, the (left) coset of W containing e_n is written as $e_n + W$ rather than $e_n W$.)

In the same way we see that an error in the ith digit converts the subgroup W of codewords into the coset $e_i + W$ where e_i is the word of length n which has all entries '0' except the ith, which is '1'. Similarly two, or more, errors result in the set of codewords being replaced by a coset of itself. For instance if $n = 10$ then an error in the third digit combined with an error in the fifth digit transforms the codeword w into the word $0010100000 + w$, so replaces the subgroup W of codewords with the coset $0010100000 + W$.

Suppose then that we receive a word which is not a codeword. We know that at least one error has occurred: we wish to recover the word that was sent without having the message retransmitted. Of course there is a problem here: even if the word received is a codeword it is possible that it is not the original word sent (for example if the distance between two codewords is three, then three errors could conspire to convert the one to the other). So we must simply be content to recover the word most likely to have been sent. This is known as *maximum likelihood decoding*. What we do therefore is to 'correct' the message by replacing the received word v by the codeword to which it is closest. Thus we compute the distances between the received word v and the various codewords w, look for the minimum distance $d(v, w)$, and replace v by w. In the event of a tie we just choose any one of the closest codewords.

Rather than do the above computation every time we receive an erroneous word, it is as well to do the computations once and for all and to prepare a table showing the result of these computations. We will now describe how to draw up such a **coset decoding table**.

We are supposing that we have a coding function $f : \mathbf{B}^m \longrightarrow \mathbf{B}^n$ for which the associated code is a linear code and we define W to be the subgroup of $\mathbf{B}^n$ consisting of all codewords. List the elements of W in some fixed order with the zero n-tuple as the first element, then array these as the top row of the decoding table. Now look for a word of $\mathbf{B}^n$ of minimum weight among those *not* in W: there will probably be more than one of these, so just choose any one, v say. Beneath each codeword w on the top row place the word $v + w$. Thus the second row of our table lists the elements of the coset $v + W$. Next look for an element u of minimum weight which is *not already listed*

in our table and list its coset, just as we listed the coset of v, to form the next line of the table (so beneath each word w of W we now have $v + w$ and, beneath that, $u + w$). Repeat this procedure until all elements of $\mathbf{B}^n$ have been listed.

This decoding table is used as follows: on receiving the word v we look for where it occurs in the table (looking for it in the top row first); having found it we replace it by the (code)word which lies directly above it on the top row.

Example Consider the coding function $f : \mathbf{B}^2 \longrightarrow \mathbf{B}^5$ and generator matrix G as in Example 2 on p. 239 (so f is a linear code). We saw that the codewords are

$$00000 \quad 01011 \quad 10110 \quad 11101.$$

We will retain this order and array them as the first line of the table. Next, look for a word of minimum weight not in this list. There are five words $e_1, \ldots, e_5$ of weight 1 and none of these is in this list. We just choose any one of them, say 00001. The second row of the table is formed by placing beneath each codeword w the word $00001 + w$ (which is as w but with its last digit changed):

$$\begin{array}{cccc} 00000 & 01011 & 10110 & 11101 \\ 00001 & 01010 & 10111 & 11100 \end{array}$$

Now look for a word of minimum weight not yet in the table. There are four choices $\{e_1, \ldots, e_4\}$; let us be systematic and take 00010 to obtain

$$\begin{array}{cccc} 00000 & 01011 & 10110 & 11101 \\ 00001 & 01010 & 10111 & 11100 \\ 00010 & 01001 & 10100 & 11111 \end{array}$$

For our next three choices we may take e_3, e_2, and e_1:

$$\begin{array}{cccc} 00000 & 01011 & 10110 & 11101 \\ 00001 & 01010 & 10111 & 11100 \\ 00010 & 01001 & 10100 & 11111 \\ 00100 & 01111 & 10010 & 11001 \\ 01000 & 00011 & 11110 & 10101 \\ 10000 & 11011 & 00110 & 01101 \end{array}$$

Since $\mathbf{B}^5$ has $2^5 = 32$ elements and the subgroup W has $2^2 = 4$ elements, there remain two rows to be added before every element of $\mathbf{B}^5$ is listed (making

eight rows in all). Every element of $\mathbf{B}^5$ of weight 1 is now in the table (as well as some others of higher weight), so in our search for elements not in the table we must start looking among those of weight 2. A number of these are already in the table but, for example, 10001 is not:

00000	01011	10110	11101
00001	01010	10111	11100
00010	01001	10100	11111
00100	01111	10010	11001
01000	00011	11110	10101
10000	11011	00110	01101
10001	11010	00111	01100

Searching among words of length two we see that 00101 has not yet appeared, so its coset gives us the last row of the table:

00000	01011	10110	11101
00001	01010	10111	11100
00010	01001	10100	11111
00100	01111	10010	11001
01000	00011	11110	10101
10000	11011	00110	01101
10001	11010	00111	01100
00101	01110	10011	11000

So if a word is transmitted with one error it will appear in the second to sixth rows. If a word is transmitted with two errors then it may appear in any row but the top one (it need not appear in one of the last two rows since two errors may bring it within distance 1 of a codeword).

How do we use this table? Suppose that the message

00 01 01 00 10 11 11 01 00

is to be sent. This will actually be transmitted (after applying the generator matrix) as:

00000 01011 01011 00000 10110 11101 11101 01011 00000.

Suppose that it is received (with a very high number of errors!) as:

00000 00011 01011 00000 11100 11101 10101 11101 01000.