Notes on Algebraic Structures

Peter J. Cameron

Preface

These are the notes of the second-year course Algebraic Structures I at Queen Mary, University of London, as I taught it in the second semester 2005–2006.

After a short introductory chapter consisting mainly of reminders about such topics as functions, equivalence relations, matrices, polynomials and permutations, the notes fall into two chapters, dealing with rings and groups respectively. I have chosen this order because everybody is familiar with the ring of integers and can appreciate what we are trying to do when we generalise its properties; there is no well-known group to play the same role. Fairly large parts of the two chapters (subrings/subgroups, homomorphisms, ideals/normal subgroups, Isomorphism Theorems) run parallel to each other, so the results on groups serve as revision for the results on rings. Towards the end, the two topics diverge. In ring theory, we study factorisation in integral domains, and apply it to the construction of fields; in group theory we prove Cayley's Theorem and look at some small groups.

The set text for the course is my own book *Introduction to Algebra*, Oxford University Press. I have refrained from reading the book while teaching the course, preferring to have another go at writing out this material.

According to the learning outcomes for the course, a studing passing the course is expected to be able to do the following:

- Give the following. Definitions of binary operations, associative, commutative, identity element, inverses, cancellation. Proofs of uniqueness of identity element, and of inverse.
- Explain the following. Group, order of a group, multiplication table for a group, subgroups and subgroup tests, cyclic subgroups, order of an element, Klein four group.
- Describe these examples: groups of units of rings, groups of symmetries of equilateral triangle and square.
- Define right cosets of a group and state Lagrange's Theorem. Explain normal subgroup, group homomorphism, kernel and image.

- Explain the following: ring, types of rings, subrings and subring tests, ideals, unit, zero-divisor, divisibility in integral domains, ring homomorphism, kernel and image.

Note: The pictures and information about mathematicians in these notes are taken from the St Andrews *History of Mathematics* website:
`http://www-groups.dcs.st-and.ac.uk/~history/index.html`

Peter J. Cameron
April 13, 2006

Contents

Chapter 1

Introduction

The first chapter of the notes will tell you a bit about what this subject involves, and then will go over material that you should be familliar with: sets, relations, functions; equivalence relations; matrices and polynomials; and permutations.

A couple of reminders about notation:

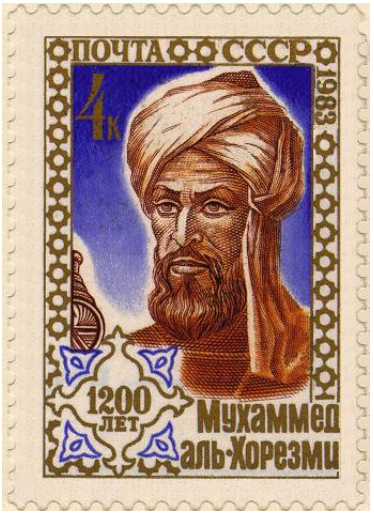
- $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ denote the natural numbers, integers, rational numbers, real numbers, and complex numbers respectively;
- if A is an $m \times n$ matrix, then A_{ij} denotes the entry in the ith row and jth column of A, for $1 \leq i \leq m$ and $1 \leq j \leq n$.

1.1 Abstract algebra

Algebra is about operations on sets. You have met many operations; for example:

- addition and multiplication of numbers;
- modular arithmetic;
- addition and multiplication of polynomials;
- addition and multiplication of matrices;
- union and intersection of sets;
- composition of permutations.

Many of these operations satisfy similar familiar laws. In all these cases, the "associative law" holds, while most (but not all!) also satisfy the "commutative law".



The name "algebra" comes from the title of the book *Hisab al-jabr w'al-muqabala* by Abu Ja'far Muhammad ibn Musa Al-Khwarizmi, a Persian mathematician who lived in Baghdad early in the Islamic era (and whose name has given us the word 'algorithm' for a procedure to carry out some operation). Al-Khwarizmi was interested in solving various algebraic equations (especially quadratics), and his method involves applying a transformation to the equation to put it into a standard form for which the solution method is known.

We will be concerned, not so much with solving particular equations, but general questions about the kinds of systems in which Al-Khwarizmi's methods might apply.

Some questions we might ask include:

(a) We form $\mathbb{C}$ by adjoining to $\mathbb{R}$ an element i satisfying $i^2 = -1$, and then assert that the "usual laws" apply in $\mathbb{C}$. How can we be sure that this is possible? What happens if we try to add more such elements?

(b) What is modular arithmetic? What exactly are the objects, and how are the operations on them defined? Does it satisfy the "usual laws"?

(c) What are polynomials? Do they satisfy the "usual laws"? What about matrices?

(d) Do union and intersection of sets behave like addition and multiplication of numbers? What about composition of permutations?

(e) What are the "usual laws"? What consequences do they have?

In this course we will define and study two kinds of algebraic object:

rings, with operations of addition and multiplication;

groups, with just one operation (like multiplication or composition).

Groups are in some ways simpler, having just a single operation, but rings are more familiar since the integers make a good prototype to think about.

1.2 Sets, functions, relations

Sets Two sets are equal if and only if they have the same members. That is,

$$A = B \quad \text{if and only if}((x \in A) \Leftrightarrow (x \in B)).$$

This means that, to prove that two sets are equal, you have to do two things:

(i) show that any element of A lies in B;

(ii) show that any element of B lies in A.

Of course, (i) means that $A \subseteq B$ (that is, A is a subset of B), while (ii) means that $B \subseteq A$. So we can re-write our rule:

$$\begin{aligned} A \subseteq B \quad &\text{if and only if} \quad ((x \in A) \Rightarrow (x \in B)), \\ A = B \quad &\text{if and only if} \quad A \subseteq B \text{ and } B \subseteq A. \end{aligned}$$

From two sets A and B we can build new ones:

union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$;

intersection: $A \cap B = \{x : x \in A \text{ and } x \in B$;

difference: $A \setminus B = \{x : x \in A \text{ and } x \notin B\}$;

symmetric difference: $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

Cartesian product

If A and B are sets, their *cartesian product* $A \times B$ is the set of all ordered pairs (a,b) for $a \in A$ and $b \in B$. The name commemorates Descartes, who showed us that we can match up the points of the Euclidean plane with $\mathbb{R} \times \mathbb{R}$ by using *cartesian coordinates*: the point x units east and y units north of the origin is matched with the pair (x,y). Then an equation connecting x and y describes the set of points on some curve in the plane: geometry meets algebra!

Functions A function f from A to B is, informally, a "black box" such that, if we input an element $a \in A$, then an element $f(a) \in B$ is output. More formally, a *function* is a set of ordered pairs (that is, a subset of the cartesian product $A \times B$) such that, for any $a \in A$, there is a unique $b \in B$ such that $(a,b) \in f$; we write $b = f(a)$ instead of $(a,b) \in f$.

The sets A and B are called the *domain* and *codomain* of f; its *image* consists of the set

$$\{b \in B : b = f(a) \text{ for some } a \in A\},$$

a subset of the codomain.

A function f is

surjective (or *onto*) if, for every $b \in B$, there is some $a \in A$ such that $b = f(a)$ (that is, the image is the whole codomain);

injective (or *one-to-one*) if $a_1 \neq a_2$ implies $f(a_1) \neq f(a_2)$ (two different elements of A cannot have the same image);

bijective if it is both injective and surjective.

Operations An operation is a special kind of function.

An *n-ary operation* on a set A is a function f from $A^n = \underbrace{A \times \cdots \times A}_{n \text{ times}}$ to A. That is, given any $a_1, \ldots, a_n \in A$, there is a unique element $b = f(a_1, \ldots, a_n) \in A$ obtained by applying the operation to these elements.

The most important cases are $n = 1$ and $n = 2$; we usually say *unary* for "1-ary", and *binary* for "2-ary". We have already seen that many binary operations (addition, multiplication, composition) occur in algebra.

Example Addition, multiplication, and subtraction are binary operations on $\mathbb{R}$, defined by

$f(a,b) = a+b$ (addition),

$f(a,b) = ab$ (multiplication),

$f(a,b) = a-b$ (subtraction).

Taking the negative is a unary operation: $f(a) = -a$.

Notation As the above example suggests, we often write binary operations, not in functional notation, but in either of two different ways:

- *infix notation*, where we put a symbol for the binary operation between the two elements that are its input, for example $a+b$, $a-b$, $a\cdot b$, $a*b$, $a\circ b$, $a\bullet b$; or
- *juxtaposition*, where we simply put the two inputs next to each other, as ab (this is most usually done for multiplication).

There are various properties that a binary relation may or may not have. Here are two. We say that the binary operation $\circ$ on A is

- *commutative* if $a\circ b = b\circ a$ for all $a,b\in A$;
- *associative* if $(a\circ b)\circ c = a\circ(b\circ c)$ for all $a,b,c\in A$.

For example, addition on $\mathbb{R}$ is commutative and associative; multiplication of 2×2 matrices is associative but not commutative; and subtraction is neither.

A binary operation $*$ on a finite set A can be represented by an *operation table*, with rows and columns labelled by elements of A. In row a and column b we put $a*b$. Here is a small example.

$*$	a	b
a	a	b
b	a	a

Relations A *binary relation* R on A is a subset of $A\times A$. If $(a,b)\in R$, we say that a and b are related, otherwise they are not related, by R.

As with operations, we often use *infix notation*, for example $a<b$, $a\leq b$, $a=b$, $a\cong b$, $a\sim b$. But note the difference:

$+$ is an operation, so $a+b$ is a member of A;

$<$ is a relation, so $a<b$ is an assertion which is either true or false.

Example Let $A=\{1,2,3\}$. Then the relation $<$ on A consists of the pairs

$$\{(1,2),(1,3),(2,3)\},$$

while the relation $\leq$ consists of the pairs

$$\{(1,1),(1,2),(1,3),(2,2),(2,3),(3,3)\}.$$

Also like operations, there are various laws or properties that a relation may have. We say that the binary operation R on A is

- *reflexive* if $(a,a) \in R$ for all $a \in A$;
- *irreflexive* if $(a,a) \notin R$ for all $a \in A$;
- *symmetric* if $(a,b) \in R$ implies $(b,a) \in R$;
- *antisymmetric* if (a,b) and (b,a) are never both in R except possibly if $a = b$;
- *transitive* if $(a,b) \in R$ and $(b,c \in R$ imply $(a,c) \in R$.

For example, $<$ is irreflexive, antisymmetric and transitive, while $\leq$ is reflexive, antisymmetric and transitive.

1.3 Equivalence relations and partitions

A binary relation R on A is an *equivalence relation* if it is reflexive, symmetric and transitive.

A *partition* P of A is a collection of subsets of A having the properties

(a) every set in P is non-empty;

(b) for every element $a \in A$, there is a *unique* set $X \in P$ such that $a \in X$.

The second condition says that the sets in P cover A without overlapping.

The first important fact we meet in the course is this:

> *Equivalence relations and partitions are essentially the same thing. Any equivalence relation on a set gives us a partition of the set, and any partition comes from a unique equivalence relation.*

We will state this as a theorem after the next definition.

Let R be an equivalence relation on a set A. For any $a \in A$, we define the *equivalence class* of a to be the set of all elements related to a: that is,

$$R(a) = \{b \in A : (a,b) \in R\}.$$

If we don't need to mention the name of the equivalence relation, we may denote the equivalence class of a by $[a]$.

Theorem 1.1 *If R is an equivalence relation on a set A, then the equivalence classes of A form a partition of A.*

Conversely, if P is a partition of A, then there is a unique equivalence relation R on A for which P is the set of equivalence classes of R.

Proof First, let R be an equivalence relation on A, and let P be the set of equivalence classes of R: that is, $P = \{R(a) : a \in A\}$. We have to show two things.

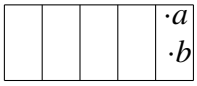
(a) First we show that the members of P are all non-empty. Take an equivalence class, say $R(a)$. By the reflexive law, $(a,a) \in R$; then, by definition, $a \in R(a)$. So $R(a)$ is not empty.

(b) Now take any element $a \in A$; we must show that a lies in exactly one equivalence class. From what we just did, we know that a lies in the equivalence class $R(a)$; so we have to show that, if a lies in another class, say, $R(b)$, then $R(b) = R(a)$. Since $a \in R(b)$, we know that $(b,a) \in R$. According to the rule for proving two sets equal, we have two things to do:

(i) Take $x \in R(a)$. By definition, $(a,x) \in R$. We know that $(b,a) \in R$. Applying the transitive law, we see that $(b,x) \in R$, that is, $x \in R(b)$. So $R(a) \subseteq R(b)$.

(ii) Take $x \in R(b)$. By definition, $(b,x) \in R$. We know that $(b,a) \in R$, and so by the symmetric law, $(a,b) \in R$. Then by the transitive law, $(a,x) \in R$, so $x \in R(a)$. Thus, $R(b) \subseteq R(a)$.

These two arguments allow us to conclude that $R(a) = R(b)$, and we are done.



If $(a,b) \in R$, then $R(a) = R(b)$.

Now we prove the converse. Suppose that P is a partition of A. We *define* a binary relation R on A by the rule that

$$(a,b) \in R \quad \text{if and only if} \quad a,b \in X \text{ for some set } X \in P.$$

We have to show that R is an equivalence relation, that is, to check the three laws.

reflexive: Take any $a \in A$. Since P is a partition, a lies in some set $X \in P$. Then $a,a \in X$, so $(a,a) \in R$ by definition.

symmetric: Suppose that $(a,b) \in R$. Then $a,b \in X$ for some $X \in P$. Since the order doesn't matter, it follows that $(b,a) \in R$.

transitive: Suppose that $(a,b) \in R$ and $(b,c) \in R$. Then there are sets $X,Y \in P$ such that $a,b \in X$, $b,c \in Y$. Now P is a partition, so b lies in a unique set of P. This means that $X = Y$. Now $a,c \in X$, so $(a,c) \in R$.

Finally we have to show that, for this relation R, the equivalence classes are the sets in P. Let $a \in A$, and let X be the unique set of P which contains a. We have to show that $X = R(a)$. As usual, there are two jobs:

if $b \in X$, then $a,b \in X$, so $(a,b) \in R$, so $b \in R(a)$.

if $b \in R(a)$, then $(a,b) \in R$, so there is some $Y \in P$ with $a,b \in Y$. But there is a unique set $X \in P$ containing a, so $Y = X$, whence $b \in X$.

Thus $X = R(a)$ as required.

Example 1 Let $f : A \to B$ be a function. Define a relation R on A by the rule that $(a_1,a_2) \in R$ if and only if $f(a_1) = f(a_2)$. Then f is an equivalence relation. (It is completely straightforward to show that R is reflexive, symmetric and transitive: try it!) There is a bijection between the equivalence classes of R and the points in the image of the function f.

For example, let $A = B = \{1,2,3,4,5\}$ and let $f(x) = x^2 - 6x + 10$. Calculation gives

x	1	2	3	4	5
$f(x)$	5	2	1	2	5

So the equivalence classes of the relation R are $\{1,5\}$, $\{2,4\}$, and $\{3\}$.

Example 2 Let n be a positive integer. Recall that two integers x and y are *congruent modulo n*, written $x \equiv y \pmod{n}$, if n divides $y - x$. This is an equivalence relation:

reflexive: n divides $0 = x - x$, so $x \equiv x \pmod{n}$.

symmetric: Suppose that $x \equiv y \pmod{n}$, so that n divides $y - x$. Then n divides $-(y-x) = x - y$, so $y \equiv x \pmod{n}$.

transitive: Suppose that $x \equiv y \pmod{n}$ and $y \equiv z \pmod{n}$. Then n divides $y - x$ and $z - y$, so divides $(y-x) + (z-y) = (z-x)$; hence $x \equiv z \pmod{n}$.

The equivalence classes of this relation are the *congruence classes modulo n*. Sometimes we write the equivalence class of x modulo n as $[x]_n$. Thus,

$$[x]_n = \{\ldots, x-2n, x-n, x, x+n, x+2n, x+3n, \ldots\},$$

an infinite set. The total number of equivalence classes is n; the classes are $[0]_n, [1]_n, \ldots, [n-1]_n$ (and then they repeat: $[n]_n = [0]_n$).

A *representative* of an equivalence class is just an element of the class. The system is completely egalitarian: anyone can be the class representative! As we have seen, if $b \in R(a)$, then $R(a) = R(b)$. A semphset of representatives for R is a set of elements, one from each equivalence class.

Sometimes there is a particularly nice way to choose the representatives, in which case they are called *canonical*. (This is not a mathematical term, since we have to decide what we mean by "nice"!) For example, the integers $\{0, 1, \ldots, n-1\}$ form a canonical set of representatives for congruence modulo n.

1.4 Matrices

You should be familiar with matrices, and with the rules for adding and multiplying them. Usually, the entries of matrices are numbers, but we will take a more general view. Thus, if S is any set, then $M_{m\times n}(S)$ means the set of all $m \times n$ matrices whose entries belong to S; a matrix is just an rectangular array with elements of S in the positions. We denote the element in row i and column j of the matrix A by A_{ij}. in the case of square matrices, with $m = n$, we simply write $M_n(S)$.

The rules for matrix operations are:

addition: if A and B are two $m \times n$ matrices, then $A + B = C$ means that $C_{ij} = A_{ij} + B_{ij}$ for $1 \le i \le m$ and $1 \le j \le n$. Note that this requires us to have a binary operation called "addition" on the set S. Note too that the two occurrences of $+$ here have different meanings: $A + B$ defines the operation of addition of matrices, while $A_{ij} + B_{ij}$ is the given operation of addition on S.

multiplication: suppose that $A \in M_{m\times n}(S)$ and $B \in M_{n\times p}(S)$, that is, the number of columns of A is the same as the number of rows of B. Then $AB = D$ means that

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

for $1 \le i \le m$, $1 \le j \le p$. This is more complicated than addition. In order to be able to multiply two matrices, we need both addition and multiplication to be defined on the elements of S. Also, we must be able to add together n

elements of S, even though $+$ is a binary operation. We will look further at how to do this later.

Remember that there are conditions on the sizes of the matrices for these rules to work: we can only add two matrices if they have the same size; and we can only multiply two matrices if the number of columns of the first is equal to the number of rows of the second. In particular, for $n \times n$ matrices, both addition and multiplication are defined.

Properties of addition and multiplication for matrices depend on properties of addition and multiplication for S.

For example, let us prove the associative law for matrix multiplication. Suppose, for convenience, that $A, B, C \in M_n(\mathbb{R})$, so that we can use all the familiar properties of addition and multiplication of real numbers. Then

$$\begin{aligned}((AB)C)_{ij} &= \sum_{k=1}^{n} (AB)_{ik} C_{kj} \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} A_{il} B_{lk} C_{kj}, \\ (A(BC))_{ij} &= \sum_{k=1}^{n} A_{ik} (BC)_{kj} \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} A_{ik} B_{kl} C_{lj}.\end{aligned}$$

These two expressions differ only by swapping the names of the "dummy variables" k and l, so are equal.

If you are not entirely comfortable with dummy variables, write out (in the case $n = 2$ the four terms in each of the two sums (with i and j fixed, but k and l each taking the values 1 and 2) and check that the results are the same.

1.5 Polynomials

Again, for the moment, the coefficients of a polynomial will be numbers; but we will generalise this later.

If you think that you know what a polynomial is, answer the following questions:

Is $1x^2 + 0x + 2$ the same polynomial as $x^2 + 2$?

Is $0x^3 + x^2 + 2$ the same polynomial as $x^2 + 2$? What is its degree?

Is $y^2 + 2$ the same polynomial as $x^2 + 2$?

I hope you answered "yes" to all these questions, and said that the degree is 2 in the second case. But these examples show that defining polynomials is going to be a bit complicated! So we defer this, and pretend for the moment that we know what a polynomial is. A first attempt at a definition would probably go: it is an expression of the form

$$\sum_{k=0}^{n} a_k x^k,$$

where we can add or delete zero terms without changing the polynomial, and x^k is short for $1x^k$. The degree of the polynomial is the exponent of the largest power of x which has a non-zero coefficient. And finally, changing the name of the variable doesn't change the polynomial.

The rules for addition and multiplication are:

addition: if $f(x) = \sum a_k x^k$ and $g(x) = \sum b_k x^k$, we can assume that both sums run from 0 to n (by adding some zero terms to one polynomial if necessary); then

$$f(x) + g(x) = \sum_{k=0}^{n} (a_k + b_k) x^k.$$

multiplication:

$$\left(\sum_{k=0}^{n} a_k x^k \right) \left(\sum_{k=0}^{m} b_k x^k \right) = \sum_{k=0}^{n+m} d_k x^k,$$

with

$$d_k = \sum_{l=0}^{k} a_l b_{k-l},$$

where any terms in this sum whose subscripts are outside the correct range are taken to be zero.

If this seems complicated, the rules are simply capturing the usual way of adding and multiplying polynomials. There is nothing mysterious here!

Addition and multiplication of real polynomials are commutative and associative operations.

1.6 Permutations

Let X be any set. A *permutation* of X is a function $g : X \to X$ which is one-to-one and onto, that is, a bijection from X to X.

There are several common notations for a permutation of the set $\{1, \ldots, n\}$. We illustrate these with the permutation of $\{1,2,3,4,5,6\}$ which maps $1 \to 3$, $2 \to 4$, $3 \to 5$, $4 \to 2$, $5 \to 1$ and $6 \to 6$.

Two-line notation: We write the numbers $1,\ldots,n$ in a row, and under each number we put its image under the permutation. In the example, this gives

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 2 & 1 & 6 \end{pmatrix}.$$

One-line notation: We write just the second line of the two-line form. In the example, this would be $(3\quad 4\quad 5\quad 2\quad 1\quad 6)$.

Cycle notation: We take the first point of the set, and follow what happens to it as we apply the permutation repeatedly. Eventually we return to the starting point. When this happens, we write the points and its images in a bracket, representing the cycle. If not every point is included, we repeat with a new point and produce another cycle, until no points are left. A point which is fixed by the permutation (mapped to itself) lies in a cycle of size 1; sometimes we don't write such cycles. In our example, this would give $(1,3,5)(2,4)(6)$, or just $(1,3,5)(2,4)$ if we choose to omit the cycle (6).

Let S_n be the set of all permutations of the set $\{1,\ldots,n\}$. We have

$$|S_n| = n! = n(n-1)(n-2)\cdots 1.$$

For consider the two-line representation. The top row is $(1\,2\,\ldots\,n)$. The bottom row consists of the same numbers in any order. Thus there are n possibilities for the first entry in the bottom row; $n-1$ possibilities for the second (anything except the first), $n-2$ possibilities for the third; and so on.

Now we define an operation on permutations as follows. If g is a permutation, denote the image of the element $x \in \{1,\ldots,n\}$ by xg. (**Warning:** we write the function on the right of its input. That is, xg, not $g(x)$ as you might expect.) Now if g and h are two permutations, their composition g_1g_2 is defined by

$$x(gh) = (xg)h \text{ for all } x \in \{1,\ldots,n\}.$$

In other words the rule is "apply g, then h".

For example, if g is the permutation $(1,3,5)(2,4)(6)$ in our above example, and $h = (1,2,3,4,5,6)$, then $gh = (1,4,3,6)(2,5)$. You are strongly urged to practice composing permutations given in cycle form!

Chapter 2

Rings

2.1 Introduction

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

2.1.1 Definition of a ring

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

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

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

Axioms for addition:

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

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

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

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

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

Axioms for multiplication:

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

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

Mixed axiom:

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

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

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

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

Further multiplicative properties

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

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

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

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

2.1.2 Examples of rings

1. The integers

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

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

2. Other number systems

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

3. The quaternions

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

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

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

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

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

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

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

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

4. Matrix rings

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

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

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

Now we have the following result:

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

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

5. Polynomial rings

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

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

where

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

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

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

Again we defer looking at the proof.

6. Rings of sets

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

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

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

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

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

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

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

7. Zero rings

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

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

8. Direct sum

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

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

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

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

The proof is straightforward checking.

9. Modular arithmetic

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

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

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

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

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

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

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

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

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

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

10. Rings of functions

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

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

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

These rings are the subject-matter of Functional Analysis.

2.1.3 Properties of rings

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

Uniqueness of zero element

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

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

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

Uniqueness of additive inverse

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

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

where we use the associative law in the third step.

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

Adding more than two elements

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

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

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

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

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

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

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

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

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

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

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

and the second as

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

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

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

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

Cancellation laws

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

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

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

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

These facts are the *cancellation laws*.

A property of zero

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

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

Commutativity of addition

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

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

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

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

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

Boolean rings

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

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

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

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

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

$$xy+yx=0.$$

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

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

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

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

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

2.1.4 Matrix rings

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

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

So we are in the position to prove Proposition 2.1.

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

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

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

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

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

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

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

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

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

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

while the second is

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

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

The proof of the other distributive law is similar.

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

2.1.5 Polynomial rings

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

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

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

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

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

Now the rules for addition and multiplication are

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

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

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

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

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

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

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

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

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

2.2 Subrings

2.2.1 Definition and test

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

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

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

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

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

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

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

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

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

We state this as a theorem:

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

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

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

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

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

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

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

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

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

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

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

2.2.2 Cosets

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

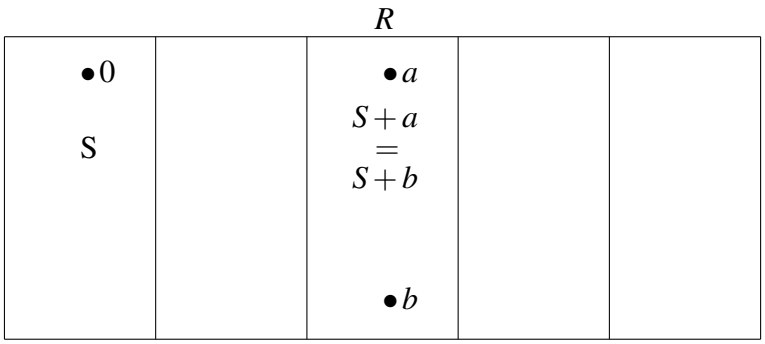
First take $b \in [a]$, so that $b-a \in S$. Let $s = b-a$. Then $b = s+a \in S+a$.

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

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

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

Here is a picture.



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

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

2.3 Homomorphisms and quotient rings

2.3.1 Isomorphism

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

$+$	o	i
o	o	i
i	i	o

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

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

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

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

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

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

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

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

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

Then θ is an isomorphism.

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

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

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

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

2.3.2 Homomorphisms

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

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

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

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

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

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

$+$	even	odd
even	even	odd
odd	odd	even

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

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

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

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

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

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

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

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

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

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

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

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

Proof We have

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

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

Then

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

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

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

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

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

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

Proof We use the Second Subring Test.

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

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

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

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

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

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

2.3.3 Ideals

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3.4 Quotient rings

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3.5 The Isomorphism Theorems

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

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

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

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

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

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

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

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

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

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

as required. So indeed ϕ is well defined.

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

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

Finally,

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

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

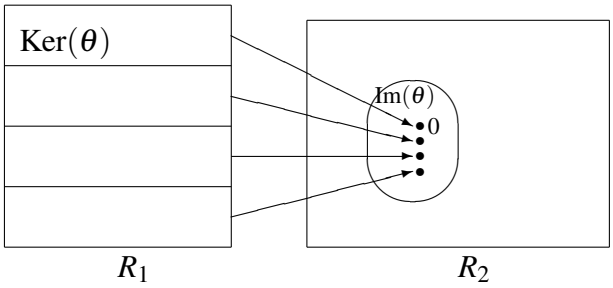


Figure 2.1: A homomorphism

We illustrate this theorem with a picture.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(c) Now the first isomorphism theorem shows that

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

and we are done.

2.4 Factorisation

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

2.4.1 Zero divisors and units

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

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

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

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

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

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

contradicting our assumption that $b \neq 0$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(a) 1 *is a unit;*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Being associates is an equivalence relation: it is

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

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

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

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

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

Now we can define greatest common divisors properly.

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

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

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

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

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

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

We abbreviate "greatest common divisor" to gcd.

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

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

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

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

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

Thus d' is a gcd of a and b.

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

2.4.2 Unique factorisation domains

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

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

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

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

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

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

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

- every element $a \in R$ which is not zero or a unit can be written as a product of irreducibles;
- if $p_1, \ldots, p_m, q_1, \ldots, q_n$ are irreducibles and

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

then $m = n$ and, after possibly permuting the factors in one product, p_i and q_i are associates for $i = 1, \ldots, m$.

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

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

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

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

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

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

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

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

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

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

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

Taking the modulus and squaring gives

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

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

Now we show that 2 is irreducible. Suppose that

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

Taking the modulus squared again gives

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

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

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

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

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

Now consider the factorisations

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

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

2.4.3 Principal ideal domains

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\langle a_1, \ldots, a_n\rangle = \{r_1a_1 + \cdots + r_na_n : r_1, \ldots, r_n \in R\}.$$

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

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

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

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

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

The claim is proved.

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

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

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

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

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

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

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

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

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

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

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

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

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

2.4.4 Euclidean domains

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

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

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

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

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

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

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

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

Both of these examples are Euclidean functions.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Theorem 2.29 *(a) A Euclidean domain is a principal ideal domain.*

(b) A Euclidean domain is a unique factorisation domain.

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

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

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

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

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

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

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

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

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

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

2.4.5 Appendix

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

The connection between the three types of domain is:

Theorem 2.31

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

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

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

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

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

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

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

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

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

2.5 Fields

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

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

A field must have more than one element.

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

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

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

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

2.5.1 Maximal ideals

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.5.2 Adding the root of a polynomial

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.5.3 Finite fields

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

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

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

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

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

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

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

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

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

We have, for example,

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

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

Galois proved the following theorem:

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

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

2.5.4 Field of fractions

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

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

(a) R is a subring of F;

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

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

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

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

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

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

We have to show that this is an equivalence relation.

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

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

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

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

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

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

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

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

a simple exercise. The proof for multiplication is similar.

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

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

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

2.5.5 Appendix: Simple rings

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

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

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

2.5.6 Appendix: The number systems

This section is not part of the course and is just for general interest. How do we build the number systems $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$?

I'll leave out the construction of $\mathbb{Z}$. Kronecker said, "God made the integers; the rest is the work of man", and though we do now know how to construct the integers (starting with nothing but the empty set), it is not straightforward.

We construct $\mathbb{Q}$ as the field of fractions of $\mathbb{Z}$.

To construct $\mathbb{R}$ from $\mathbb{Q}$, we borrow an idea from analysis, the definition of a *Cauchy sequence*: the sequence $(a_0, a_1, a_2, \ldots)$ is a Cauchy sequence if, given any $\varepsilon > 0$, there exists a positive integer N such that, for all $m, n > N$, we have $|a_m - a_n| < \varepsilon$.

We let R be the set of all Cauchy sequences of rational numbers. We make R into a ring by adding and multiplying sequences term by term. Then let I be the set of all null sequences of rational numbers (sequences which converge to 0.) Then it can be shown that R is a commutative ring with identity, and I a maximal ideal; so R/I is a field. This is the field $\mathbb{R}$ of real numbers.

We saw that $\mathbb{C}$ is constructed from $\mathbb{R}$ by adding a root of the irreducible polynomial $x^2 + 1$: that is, $\mathbb{C} = \mathbb{R}[x]/\langle x^2 + 1 \rangle$.

Chapter 3

Groups

In the remainder of the notes we will be talking about groups. A group is a structure with just one binary operation, satisfying four axioms. So groups are only half as complicated as rings! As well as being new material, this part will help you revise the first part of the course, since a lot of things (subgroups, homomorphisms, Isomorphism Theorems) work in almost exactly the same way as for rings.

3.1 Introduction

3.1.1 Definition of a group

A group is a set G with one binary operation (which we write for now as $\circ$ in infix notation[1]) satisfying the following four axioms (G0)–(G3):

(G0) (*Closure law*) For any $g,h \in G$, we have $g \circ h \in G$.

(G1) (*Associative law*) For any $g,h,k \in G$, we have $(g \circ h) \circ k = g \circ (h \circ k)$.

(G2) (*Identity law*) There is an element $e \in G$ with the property that $g \circ e = e \circ g = g$ for all $g \in G$. (The element e is called the *identity element* of G.)

(G3) (*Inverse law*) For any element $g \in G$, there is an element $h \in G$ satisfying $g \circ h = h \circ g = e$. (We denote this element h by g^{-1}, and call it the *inverse* of g.)

If a group G also satisfies the condition

[1]Remember that this means that the result of applying the operation to a and b is written as $a \circ b$.

(G4) (*Commutative law*) For any $g, h \in G$, we have $g \circ h = h \circ g$,

then G is called a *commutative group* or (more often) an *Abelian group*.

3.1.2 Examples of groups

Axioms (G0)–(G4) for a group are just axioms (A0)–(A4) for a ring but using slightly different notation (the set is G instead of R, the operation is $\circ$ instead of $+$, and so on). So we get our first class of examples:

Proposition 3.1 *Let R be a ring. Then R with the operation of addition is an Abelian group: the identity element is 0, and the inverse of a is $-a$.*

This group is called the *additive group* or R.
This is not the only way to get groups from rings.

Proposition 3.2 *Let R be a ring with identity, and $U(R)$ the set of units of R. Then $U(R)$, with the operation of multiplication, is a group. If R is a commutative ring, then $U(R)$ is an Abelian group.*

This group is called the *group of units* of R.

Proof Look back to Lemma 2.20. Let $U(R)$ be the set of units of R.

(G0) Lemma 5.1(c) shows that, if u and v are units, then so is uv. So $U(R)$ is closed for multiplication.

(G1) The associative law for multiplication holds for all elements of R (by Axiom (M1) for rings), and so in particular for units.

(G2) Lemma 5.1(a) shows that 1 is a unit. It clearly plays the role of the identity element.

(G3) Lemma 5.1(b) shows that, if u is a unit, then so is u^{-1}.

(G4) For the last part of the Proposition, if R is a commutative ring, then (M4) holds, so that $uv = vu$ for all $u, v \in R$; in particular, this holds when u and v are units.

Example If F is a field, then every non-zero element of F is a unit. So the set of non-zero elements forms an Abelian group with the operation of multiplication. This is called the *multiplicative group* of the field.

Example A matrix A is a unit in $M_n(F)$, where F is a field, if and only if $\det(A) \neq 0$. So the set of matrices with non-zero determinant is a group. This group is called the *general linear group*, and written $\mathrm{GL}(n,F)$. If $n > 1$, this group is not Abelian.

Direct product This construction corresponds to the direct sum for rings.

Let G_1 and G_2 be groups. The *direct product* $G_1 \times G_2$ is defined as follows:

- the set of elements is the Cartesian product (which is also denoted by $G_1 \times G_2$), the set of all ordered pairs (g_1, g_2) with $g_1 \in G_1$ and $g_2 \in G_2$;
- the group operation is "componentwise", that is,

$$(g_1, g_2) \circ (h_1, h_2) = (g_1 \circ h_1, g_2 \circ h_2).$$

It is an exercise to prove that it is a group.

Cayley tables As with any binary operation, the group operation can be represented by giving an operation table. In the case of a group, the operation table is usually called the *Cayley table* of the group. In principle, given the Cayley table of a finite group, we could check that all the group axioms are satisfied.

Here, for example, are the Cayley tables of two groups each with four elements.

$\circ$	e	x	y	z
e	e	x	y	z
x	x	y	z	e
y	y	z	e	x
z	z	e	x	y

$\circ$	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

Each group is Abelian, as we can see because the tables are symmetric. These two groups are obviously different: in the second group, each element is equal to its inverse, whereas this is not true in the first group. (When we come to define isomorphism, we will say that the two groups are not isomorphic.)

In fact, these groups are the additive groups of the rings $\mathbb{Z}/4\mathbb{Z}$ and $\mathrm{GF}(4)$ respectively.

3.1.3 Properties of groups

Some of these properties will look very familiar, since they are similar to what we saw for rings.

Uniqueness of identity element

The identity element of a group is unique. For suppose that there are two identity elements, say e_1 and e_2. (This means that $g \circ e_1 = e_1 \circ g = g$ for all g, and also $g \circ e_2 = e_2 \circ g = g$ for all g.) Then

$$e_1 = e_1 \circ e_2 = e_2.$$

Uniqueness of inverse

The inverse of a group element g is unique. For suppose that h and k are both additive inverses of g. (This means that $g \circ h = h \circ g = e$ and $g \circ k = k \circ g = e$ – we know now that there is a unique identity element e). Then

$$h = h \circ e = h \circ (g \circ k) = (h \circ g) \circ k = e \circ k = k,$$

where we use the associative law in the third step.

We denote the inverse of g by g^{-1}.

Composing more than two elements

We showed in Proposition 2.6 that, as long as the associative law holds, the result of composing any number of elements is independent of the way that the product is bracketed: for example, $a \circ ((b \circ c) \circ d) = (a \circ b) \circ (c \circ d)$. Since the associative law holds in a group, we have:

Proposition 3.3 *Let $g_1, \ldots, g_n$ be elements of a group G. Then the composition*

$$g_1 \circ g_2 \circ \cdots \circ g_n$$

is well-defined, independent of the way it is bracketed.

Cancellation laws

Proposition 3.4 *In a group G, if $a \circ g = b \circ g$, then $a = b$. Similarly, if $g \circ a = g \circ b$, then $a = b$.*

Proof Suppose that $a \circ g = b \circ g$, and let $h = g^{-1}$. Then

$$a = a \circ e = a \circ (g \circ h) = (a \circ g) \circ h = (b \circ g) \circ h = b \circ (g \circ h) = b \circ e = b.$$

The other law is proved similarly.

These facts are the *cancellation laws*.

Proposition 3.5 *The inverse of $g \circ h$ is $h^{-1} \circ g^{-1}$.*

Proof

$$(g \circ h) \circ (h^{-1} \circ g^{-1}) = g \circ (h \circ h^{-1}) \circ g^{-1} = g \circ e \circ g^{-1} = g \circ g^{-1} = e,$$

using the associative law; and similarly $(h^{-1} \circ g^{-1}) \circ (g \circ h) = e$.

3.1.4 Notation

The notation $g \circ h$ for the group operation is a bit cumbersome, and we now change things.

If we are only interested in Abelian groups, we use $+$ as the symbol for the group operation, 0 for the group identity, and $-g$ for the inverse of g. This agrees with the additive notation in a ring. Indeed, the additive group of a ring is an Abelian group, and every Abelian group is the additive group of a ring. [To see this, take the group operation as addition, and construct the zero ring: all products are zero.]

For general groups which may not be Abelian, we use juxtaposition for the group operation, 1 for the identity, and g^{-1} for the inverse of g. (This is like multiplicative notation in a ring, but it is not true that every group is the group of units in some ring!!)

This table gives the correspondences.

Type of group	Operation	Notation	Identity	Inverse
General	$\circ$	$g \circ h$	e	g^{-1}
Abelian	$+$	$g+h$	0	$-g$
General	Juxtaposition	gh	1	g^{-1}

For the rest of this course, our notation for the group operation will be juxtaposition.

3.1.5 Order

The term *order* has two quite different meanings in group theory: be careful not to confuse them. In the next chapter we will see that there is a close relationship between the two meanings.

The *order* of a group is the number of elements of the group. It may be finite (in which case it is a positive integer), or infinite.

To define the second kind of order, we introduce the notation g^n. This means the result of composing n factors g together:

$$g^n = gg\cdots g \text{ (n factors).}$$

More formally, $g^0 = 1$, and for any positive integer n, $g^n = g \cdot g^{n-1}$.

The *order* of an element g in a group is defined as follows:

- If $g^n = 1$ for some positive integer n, then the smallest such n is called the order of g.
- If no such n exists, we say that g has infinite order.

Thus, the identity element always has order 1. If an element g has order 2, then it is equal to its inverse (for $g^2 = 1 = gg^{-1}$ implies $g = g^{-1}$ by the Cancellation Law.)

Consider the additive group of the ring $\mathbb{Z}$. (Recall that the operation is $+$ and the zero element is 0; so, instead of g^n we write $n \cdot g$, and the order is the smallest positive n such that $n \cdot g = 0$, or is infinite if no such n exists.) The element 1 has infinite order, since there is no positive integer n such that $n \cdot 1 = 0$.

In the first group in our two examples above of Cayley tables, the elements x and z have order 4 (we have $x^2 = y$, $x^3 = z$, $x^4 = e$ which is the identity element), while y has order 2. In the second group, all of a, b, c have order 2.

3.1.6 Symmetric groups

We end this chapter by defining an important class of groups.

Let X be any set. A *permutation* of X is a function $g : X \to X$ which is one-to-one and onto, that is, a bijection from X to X. Recall the discussion of permutations in Chapter 1.

Let S_n be the set of all permutations of the set $\{1,\ldots,n\}$. We have

$$|S_n| = n! = n(n-1)(n-2)\cdots 1.$$

For consider the two-line representation. The top row is $(1\,2\,\ldots\,n)$. The bottom row consists of the same numbers in any order. Thus there are n possibilities for the first entry in the bottom row; $n-1$ possibilities for the second (anything except the first), $n-2$ possibilities for the third; and so on.

Now we define an operation on permutations as follows. If g is a permutation, denote the image of the element $x \in \{1,\ldots,n\}$ by xg. (As with homomorphisms, we write the function on the right of its input.) Now if g and h are two permutations, their composition g_1g_2 is defined by

$$x(gh) = (xg)h \text{ for all } x \in \{1,\ldots,n\}.$$

In other words the rule is "apply g, then h".

For example, if g is the permutation $(1,3,5)(2,4)(6)$ in our above example, and $h = (1,2,3,4,5,6)$, then $gh = (1,4,3,6)(2,5)$. You are strongly urged to practice composing permutations given in cycle form!

Theorem 3.6 *The set S_n of permutations of $\{1,\ldots,n\}$, with the operation of composition, is a group.*

Proof (G0) If g and h are bijections, we have to show that gh is a bijection.

- To show that it is one-to-one, suppose that $x(gh) = y(gh)$. By definition this means $(xg)h = (yg)h$. Since h is one-to-one, this implies $xg = yg$; then, since g is one-to-one, this implies $x = y$.
- To show that it is onto, choose any element $z \in \{1,\ldots,n\}$. Since h is onto, we can find y such that $yh = z$. Then since g is onto, we can find x such that $xg = y$. Then $x(gh) = (xg)h = yh = z$.

(G1) Let g,h,k be three permutations. To show that $g(hk) = (gh)k$, we have to show that these two permutations have the same effect on any element $x \in \{1,\ldots,n\}$. Now we have

$$x(g(hk)) = (xg)(hk) = ((xg)h)k) = (x(gh))k = x((gh)k),$$

as required.

(G2) The identity permutation 1 is the permutation which leaves everything as it was: that is, $x1 = x$ for all $x \in \{1,\ldots,n\}$. Then $x(1g) = (x1)g = xg$ for all x, so that $1g = g$; similarly $g1 = g$.

(G3) The inverse of a permutation g is simply the "inverse function" which undoes the effect of g: that is, $xg^{-1} = y$ if $yg = x$. Then it is clear that $gg^{-1} = g^{-1}g = 1$.

We call this group the *symmetric group* on n symbols, and denote it by S_n.

Proposition 3.7 *S_n is not Abelian for $n \geq 3$.*

Proof If $g = (1,2)$ and $h = (1,3)$ (all other points are fixed), then $gh = (1,2,3)$ but $hg = (1,3,2)$.

Exercise Show that S_2 is an Abelian group.

Exercise Verify the following Cayley table for S_3:

	1	$(1,2,3)$	$(1,3,2)$	$(1,2)$	$(2,3)$	$(1,3)$
1	1	$(1,2,3)$	$(1,3,2)$	$(1,2)$	$(2,3)$	$(1,3)$
$(1,2,3)$	$(1,2,3)$	$(1,3,2)$	1	$(2,3)$	$(1,3)$	$(1,2)$
$(1,3,2)$	$(1,3,2)$	1	$(1,2,3)$	$(1,3)$	$(1,2)$	$(2,3)$
$(1,2)$	$(1,2)$	$(1,3)$	$(2,3)$	1	$(1,3,2)$	$(1,2,3)$
$(2,3)$	$(2,3)$	$(1,2)$	$(1,3)$	$(1,2,3)$	1	$(1,3,2)$
$(1,3)$	$(1,3)$	$(2,3)$	$(1,2)$	$(1,3,2)$	$(1,2,3)$	1

We end with a result which you probably met in Discrete Maths.

Proposition 3.8 *Let g be an element of S_n, written in cycle notation. Then the order of g is the least common multiple of its cycle lengths.*

Proof Take any cycle of g, say of length k. Then the points in this cycle return to their original position after g has been applied k times. So g^n fixes the points of this cycle if and only if n is a multiple of k.

We deduce that $g^n = 1$ if and only if n is a multiple of every cycle length. So the order of g is the least common multiple of the cycle lengths.

So, of the elements of S_3, the identity has order 1, the elements $(1,2)$, $(2,3)$ and $(1,3)$ have order 2, and $(1,2,3)$ and $(1,3,2)$ have order 3. (Remember that $(1,2)$ is really $(1,2)(3)$, with a cycle of length 1; but this doesn't alter the least common multiple of the cycle lengths.)

3.2 Subgroups

This section corresponds to Section 2.2 on subrings.

3.2.1 Subgroups and subgroup tests

A *subgroup* of a group G is a subset of G which is a subgroup in its own right (with the same group operation).

There are two subgroup tests, resembling the two subring tests.

Proposition 3.9 (First Subgroup Test) *A non-empty subset H of a group G is a subgroup of G if, for any $h,k \in H$, we have $hk \in H$ and $h^{-1} \in H$.*

Proof We have to show that H satisfies the group axioms. The conditions of the test show that it is closed under composition (G0) and inverses (G3). The associative law (G1) holds in H because it holds for all elements of G. We have only to prove (G2), the identity axiom.

We are given that H is non-empty, so choose $h \in H$. Then by assumption, $h^{-1} \in H$, and then (choosing $k = h^{-1}$) $1 = hh^{-1} \in H$.

We can reduce the number of things to be checked from two to one:

Proposition 3.10 (Second Subgroup Test) *A non-empty subset H of a group G is a subgroup of G if, for any $h,k \in H$, we have $hk^{-1} \in H$.*

Proof Choosing $k = h$, we see that $1 = hh^{-1} \in H$. Now using 1 and h in place of h and k, we see that $h^{-1} = 1h^{-1} \in H$. Finally, given $h,k \in H$, we know that $k^{-1} \in H$, so $hk = h(k^{-1})^{-1} \in H$. So the conditions of the First Subgroup Test hold.

Example Look back to the Cayley tables in the last chapter. In the first case, $\{e,y\}$ is a subgroup. In the second case, $\{e,a\}$, $\{e,b\}$ and $\{e,c\}$ are all subgroups.

3.2.2 Cyclic groups

If g is an element of a group G, we define the powers g^n of G (for $n \in \mathbb{Z}$) as follows: if n is positive, then g^n is the product of n factors g; $g^0 = 1$; and $g^{-n} = (g^{-1})^n$. The usual laws of exponents hold: $g^{m+n} = g^m \cdot g^n$ and $g^{mn} = (g^m)^n$.

A *cyclic group* is a group C which consists of all the powers (positive and negative) of a single element. If C consists of all the powers of g, then we write $C = \langle g \rangle$, and say that C is *generated by* g.

Proposition 3.11 *A cyclic group is Abelian.*

Proof Let $C = \langle g \rangle$. Take two elements of C, say g^m and g^n. Then

$$g^m \cdot g^n = g^{m+n} = g^n \cdot g^m.$$

Let $C = \langle g \rangle$. Recall the *order* of g, the smallest positive integer n such that $g^n = 1$ (if such n exists – otherwise the order is infinite).

Proposition 3.12 *Let g be an element of the a group G. Then the set of all powers (positive and negative) of g forms a cyclic subgroup of G. Its order is equal to the order of g.*

Proof Let $C = \{g^n : n \in \mathbb{Z}\}$. We apply the Second Subgroup test: if $g^m, g^n \in C$, then $(g^m)(g^n)^{-1} = g^{m-n} \in C$. So C is a subgroup.

If g has infinite order, then no positive power of g is equal to 1. It follows that all the powers g^n for $n \in \mathbb{Z}$ are different elements. (For if $g^m = g^n$, with $m > n$, then $g^{n-m} = 1$.) So C is infinite.

Suppose that g has finite order n. We claim that any power of g is equal to one of the elements $g^0 = 1, g^1 = g, \ldots, g^{n-1}$. Take any power g^m. Using the division algorithm in $\mathbb{Z}$, write $m = nq + r$, where $0 \leq r \leq n-1$. Then

$$g^m = g^{nq+r} = (g^n)^q \cdot g^r = 1 \cdot g^r = g^r.$$

Furthermore, the elements $1, g, \ldots, g^{n-1}$ are all different; for if $g^r = g^s$, with $0 \leq r < s \leq n-1$, then $g^{s-r} = 1$, and $0 < s-r < n$, contradicting the fact that n is the order of g (the smallest exponent i such that $g^i = 1$).

Example The additive group of the ring $\mathbb{Z}/n\mathbb{Z}$ is a cyclic group of order n, generated by $\bar{1} = n\mathbb{Z} + 1$. Remember that the group operation is addition here, and the identity element is zero, so in place of $g^n = 1$ we have $n\bar{1} = \bar{0}$, which is true in the integers mod n; moreover it is true that no smaller positive multiple of $\bar{1}$ can be zero.

Proposition 3.13 *Let G be a cyclic group of finite order n. Then g has a cyclic subgroup of order m for every m which divides n; and these are all the subgroups of G.*

Proof Let $G = \langle g \rangle = \{1, g, g^2, \ldots, g^{n-1}\}$. If m divides n, let $n = mk$, and put $h = g^k$. Then $h^m = (g^k)^m = g^n = 1$, and clearly no smaller power of h is equal to 1; so h has order m, and generates a cyclic group of order m.

Now let H be any subgroup of G. If $H = \{1\}$, then H is the unique cyclic subgroup of order 1 in G, so suppose not. Let g^m be the smallest positive power of g which belongs to H. We claim that, if $g^k \in H$, then m divides k. For let $k = mq + r$, where $0 \leq r \leq m-1$. Then

$$g^r = g^{mq+r} g^{-mq} = g^k (g^m)^{-q} \in H,$$

so $r = 0$ (since m was the smallest positive exponent of an element of H. So H is generated by g^m. Now $g^n = 1 \in H$, so m divides n, and we are done.

3.2.3 Cosets

Given any subgroup H of a group G, we can construct a partition of G into "cosets" of H, just as we did for rings. But for groups, things are a bit more complicated.

Because the group operation may not be commutative, we have to define two different sorts of cosets.

Let H be a subgroup of a group G. Define a relation $\sim_r$ on G by the rule

$$x \sim_r y \text{ if and only if } yx^{-1} \in H.$$

We claim that $\sim_r$ is an equivalence relation:

reflexive: For any $x \in G$, we have $xx^{-1} = 1 \in H$, so $x \sim_r x$.

symmetric: Suppose that $x \sim_r y$, so that $h = yx^{-1} \in H$. Then $h^{-1} = (yx^{-1})^{-1} = xy^{-1} \in H$, so $y \sim_r x$.

transitive: Suppose that $x \sim_r y$ and $y \sim_r z$, so that $h = yx^{-1} \in H$ and $k = zy^{-1} \in H$. Then $kh = (zy^{-1})(yx^{-1}) = zx^{-1} \in H$, so $x \sim_r z$.

The equivalence classes of this equivalence relation are called the *right cosets* of H in G.

A right coset is a set of elements of the form $Hx = \{hx : h \in H\}$, for some fixed element $x \in G$ called the "coset representative". For

$$y \in Hx \Leftrightarrow y = hx \text{ for some } h \in H \leftrightarrow yx^{-1} \in H \Leftrightarrow x \sim_r y.$$

We summarise all this as follows:

Proposition 3.14 *If H is a subgroup of the group G, then G is partitioned into right cosets of H in G, sets of the form $Hx = \{hx : h \in H\}$.*

In a similar way, the relation $\sim_l$ defined on G by the rule

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

is an equivalence relation on G, and its equivalence classes are the *left cosets* of H in G, the sets of the form $xH = \{xh : h \in H\}$.

If G is an abelian group, the left and right cosets of any subgroup coincide, since

$$Hx = \{hx : h \in H\} = \{xh : h \in H\} = xH.$$

This is not true in general:

Example Let G be the symmetric group S_3, and let H be the subgroup $\{1,(1,2)\}$ consisting of all permutations fixing the point 3. The right cosets of H in G are

$$\begin{aligned} H1 &= \{1,(1,2)\}, \\ H(1,3) &= \{(1,3),(1,2,3)\}, \\ H(2,3) &= \{(2,3),(1,3,2), \end{aligned}$$

while the left cosets are

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

We see that, as expected, both right and left cosets partition G, but the two partitions are not the same. But each partition divides G into three sets of size 2.

3.2.4 Lagrange's Theorem

Lagrange's Theorem states a very important relation between the orders of a finite group and any subgroup.

Theorem 3.15 (Lagrange's Theorem) *Let H be a subgroup of a finite group G. Then the order of H divides the order of G.*

Proof We already know from the last section that the group G is partitioned into the right cosets of H. We show that every right coset Hg contains the same number of elements as H.

To prove this, we construct a bijection ϕ from H to Hg. The bijection is defined in the obvious way: ϕ maps h to hg.

- ϕ is one-to-one: suppose that $\phi(h_1) = \phi)h_2$, that is, $h_1g = h_2g$. Cancelling the g (by the cancellation law, or by multiplying by g^{-1}), we get $h_1 = h_2$.

- ϕ is onto: by definition, every element in the coset Hg has the form hg for some $h \in H$, that is, it is $\phi(h)$.

So ϕ is a bijection, and $|Hg| = |H|$.

Now, if m is the number of right cosets of H in G, then $m|H| = |G|$, so $|H|$ divides $|G|$.

Remark We see that $|G|/|H|$ is the number of right cosets of H in G. This number is called the *index* of H in G.

We could have used left cosets instead, and we see that $|G|/|H|$ is also the number of left cosets. So these numbers are the same. In fact, there is another reason for this:

Exercise Show that the set of all inverses of the elements in the right coset Hg form the left coset $g^{-1}H$. So there is a bijection between the set of right cosets and the set of left cosets of H.

In the example in the preceding section, we had a group S_3 with a subgroup having three right cosets and three left cosets; that is, a subgroup with index 3.

Corollary 3.16 *let g be an element of the finite group G. Then the order of g divides the order of G.*

Proof Remember, first, that the word "order" here has two quite different meanings: the order of a group is the number of elements it has; while the order of an element is the smallest n such that $g^n = 1$.

However, we also saw that if the element g has order m, then the set $\{1, g, g^2, \ldots, g^{m-1}\}$ is a cyclic subgroup of G having order m. So, by Lagrange's Theorem, m divides the order of G.

Example Let $G = S_3$. Then the order of G is 6. The element $(1)(2,3)$ has order 2, while the element $(1,3,2)$ has order 3.

3.3 Homomorphisms and normal subgroups

This section is similar to the corresponding one for rings. Homomorphisms are maps preserving the structure, while normal subgroups do the same job for groups as ideals do for rings: that is, they are kernels of homomorphisms. The structure of this section follows closely that of Section 2.3.

3.3.1 Isomorphism

Just as for rings, we say that groups are isomorphic if there is a bijection between them which preserves the algebraic structure.

Formally, let G_1 and G_2 be groups. The map $\theta : G_1 \to G_2$ is an *isomorphism* if it is a bijection from G_1 to G_2 and satisfies

$$(gh)\theta = (g\theta)(h\theta) \text{ for all } g, h \in G_1.$$

Note that, as before, we write the map θ to the right of its argument: that is, $g\theta$ is the image of g under the map θ. If there is an isomorphism from G_1 to G_2, we say that the groups G_1 and G_1 are *isomorphic*.

Example Let G_1 be the additive group of $\mathbb{Z}/2\mathbb{Z}$, and let G_2 be the symmetric group S_2. Their Cayley tables are:

$+$	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{1}$
$\bar{1}$	$\bar{1}$	$\bar{0}$

$\cdot$	1	$(1,2)$
1	1	$(1,2)$
$(1,2)$	$(1,2)$	1

The map θ that takes $\bar{0}$ to 1, and $\bar{1}$ to $(1,2)$, is clearly an isomorphism from G_1 to G_2.

3.3.2 Homomorphisms

An isomorphism between groups has two properties: it is a bijection; and it preserves the group operation. If we relax the first property but keep the second, we obtain a homomorphism. Just as for rings, we say that a function $\theta : G_1 \to G_2$ is

- a *homomorphism* if it satisfies

$$(gh)\theta = (g\theta)(h\theta); \tag{3.1}$$

- a *monomorphism* if it satisfies (3.1) and is one-to-one;
- an *epimorphism* if it satisfies (3.1) and is onto;
- an *isomorphism* if it satisfies (3.1) and is one-to-one and onto.

We have the following lemma, proved in much the same way as for rings:

Lemma 3.17 *Let $\theta : G_1 \to G_2$ be a homomorphism. Then $1\theta = 1$; $(g^{-1})\theta = (g\theta)^{-1}$; and $(gh^{-1})\theta = (g\theta)(h\theta)^{-1}$, for all $g,h \in G_1$.*

Now, if $\theta : G_1 \to G_2$ is a homomorphism, we define the *image* of θ to be the subset

$$\{x \in G_2 : x = g\theta \text{ for some } g \in G_1\}$$

of G_2, and the *kernel* of θ to be the subset

$$\{g \in G_1 : g\theta = 1\}$$

of G_1.

Proposition 3.18 *Let $\theta : G_1 \to G_2$ be a homomorphism.*

(a) $\mathrm{Im}(\theta)$ *is a subgroup of* G_2.

(b) $\mathrm{Ker}(\theta)$ *is a subgroup of* G_1.

Proof We use the Second Subgroup Test in each case.

(a) Take $x, y \in \mathrm{Im}(\theta)$, say $x = g\theta$ and $y = h\theta$ for $g, h \in G_1$. Then $xy^{-1} = (gh^{-1})\theta \in \mathrm{Im}(\theta)$, by the Lemma.

(b) Take $g, h \in \mathrm{Ker}(\theta)$. Then $g\theta = h\theta = 1$, so $(gh^{-1})\theta = 1^{-1}1 = 1$; so $gh^{-1} \in \mathrm{Ker}(\theta)$.

Example Look back to the Cayley table of the symmetric group S_3 in Chapter 7. Colour the elements 1, $(1,2,3)$ and $(1,3,2)$ red, and the elements $(1,2)$, $(2,3)$ and $(1,3)$ blue. We see that the Cayley table has the "simplified form"

$\cdot$	red	blue
red	red	blue
blue	blue	red

This is a group of order 2, and the map θ taking 1, $(1,2,3)$ and $(1,3,2)$ to red and $(1,2)$, $(2,3)$ and $(1,3)$ to blue is a homomorphism. Its kernel is the subgroup $\{1,(1,2,3),(1,3,2)\}$.

3.3.3 Normal subgroups

A normal subgroup is a special kind of subgroup of a group. Recall from the last chapter taht any subgroup H has right and left cosets, which may not be the same. We say that H is a *normal subgroup* of G if the right and left cosets of H in G are the same; that is, if $Hx = xH$ for any $x \in G$.

There are several equivalent ways of saying the same thing. We define

$$x^{-1}Hx = \{x^{-1}hx : h \in H\}$$

for any element $x \in G$.

Proposition 3.19 *Let H be a subgroup of G. Then the following are equivalent:*

(a) H is a normal subgroup, that is, $Hx = xH$ for all $x \in G$;

(b) $x^{-1}Hx = H$ for all $x \in G$;

(c) $x^{-1}hx \in H$, for all $x \in G$ and $h \in H$.

Proof If $Hx = xH$, then $x^{-1}Hx = x^{-1}xH = H$, and conversely. So (a) and (b) are equivalent.

If (b) holds then every element $x^{-1}hx$ belongs to $x^{-1}Hx$, and so to H, so (c) holds. Conversely, suppose that (c) holds. Then every element of $x^{-1}Hx$ belongs to H, and we have to prove the reverse inclusion. So take $h \in H$. Putting $y = x^{-1}$, we have $k = y^{-1}hy = xhx^{-1} \in H$, so $h \in x^{-1}Hx$, finishing the proof.

Now the important thing about normal subgroups is that, like ideals, they are kernels of homomorphisms.

Proposition 3.20 *Let $\theta : G_1 \to G_2$ be a homomorphism. Then* $\mathrm{Ker}(\theta)$ *is a normal subgroup of G_1.*

Proof Let $H = \mathrm{Ker}(\theta)$. Suppose that $h \in H$ and $x \in G$. Then

$$(x^{-1}hx)\theta = (x^{-1})\theta \cdot h\theta \cdot x\theta = (x\theta)^{-1} \cdot 1 \cdot x\theta = 1,$$

so $x^{-1}hx \in \ker(\theta) = H$. By part (c) of the preceding Proposition, H is a normal subgroup of G.

There are a couple of situations in which we can guarantee that a subgroup is normal.

Proposition 3.21 *(a) If G is Abelian, then every subgroup H of G is normal.*

(b) If H has index 2 in G, then H is normal in G.

Proof (a) If G is Abelian, then $xH = Hx$ for all $x \in G$.

(b) Recall that this means that H has exactly two cosets (left or right) in G. One of these cosets is H itself; the other must consist of all the other elements of G, that is, $G \setminus H$. This is the case whether we are looking at left or right cosets. So the left and right cosets are the same.

Remark We saw in the last chapter an example of a group S_3 with a non-normal subgroup having index 3 (that is, just three cosets). So we can't improve this theorem from 2 to 3.

In our example in the last section, the subgroup $\{1, (1,2,3), (1,3,2)\}$ of S_3 has index 2, and so is normal, in S_3; this also follows from the fact that it is the kernel of a homomorphism.

For the record, here is a normal subgroup test:

Proposition 3.22 (Normal subgroup test) *A non-empty subset H of a group G is a normal subgroup of G if the following hold:*

(a) for any $h,k \in H$, we have $hk^{-1} \in H$;

(b) for any $h \in H$ and $x \in G$, we have $x^{-1}hx \in H$.

Proof (a) is the condition of the Second Subgroup Test, and we saw that (b) is a condition for a subgroup to be normal.

3.3.4 Quotient groups

Let H be a normal subgroup of a group G. We define the *quotient group* G/H as follows:

- The elements of G/H are the cosets of H in G (left or right doesn't matter, since H is normal);
- The group operation is defined by $(Hx)(hy) = Hxy$ for all $x,y \in G$; in other words, to multiply cosets, we multiply their representatives.

Proposition 3.23 *If H is a normal subgroup of G, then the quotient group G/H as defined above is a group. Moreover, the map θ from G to G/H defined by $x\theta = Hx$ is a homomorphism whose kernel is H and whose image is G/H.*

Proof First we have to show that the definition of the group operation is a good one. In other words, suppose that we chose different coset representatives x' and y' for the cosets Hx and Hy; is it true that $Hxy = Hx'y'$? We have $x' = hx$ and $y' = ky$, for some $h,k \in H$. Now xk belongs to the left coset xH. Since H is normal, this is equal to the right coset Hx, so that $xk = lx$ for some $l \in H$. Then $x'y' = hxky = (hl)(xy) \in Hxy$, since $hl \in H$. Thus the operation is indeed well-defined.

Now we have to verify the group axioms.

(G0) Closure is clear since the product of two cosets is a coset.

(G1) Given three cosets Hx, Hy, Hz, we have

$$((Hx)(Hy))(Hz) = (Hxy)(Hz) = H(xy)z = Hx(yz) = (Hx)(Hyz) = (Hx)((Hy)(Hz)),$$

using the associative law in G.

(G2) The identity is $H1 = H$, since $(H1)(Hx) = H(1x) = Hx$ for all $x \in G$.

(G3) The inverse of Hx is clearly Hx^{-1}.

Finally, for the map θ, we have

$$(xy)\theta = Hxy = (Hx)(Hy) = (x\theta)(y\theta),$$

so θ is a homomorphims. Its image consists of all cosets Hx, that is, $\mathrm{Im}(\theta) = G/H$. The identity element of G/H is (as we saw in in the proof of (G2)) the coset H; and $Hx = H$ if and only if $x \in H$, so that $\mathrm{Ker}(\theta) = H$.

The map θ in the above proof is called the *natural homomorphism* from G to G/H. We see that, if H is a normal subgroup of G, then it is the kernel of the natural homomorphism from G to G/H. So normal subgroups are the same thing as kernels of homomorphisms.

Example Let $G = S_3$, and let H be the subgroup $\{1,(1,2,3),(1,3,2)\}$. We have observed that H is a normal subgroup. It has two cosets, namely $H1 = H$ and $H(1,2) = \{(1,2),(2,3),(1,3)\}$. The rules for multiplication of these cosets will be the same as the rules for multiplying the elements 1 and $(1,2)$. So G/H is isomorphic to the group $\{1,(1,2)\}$ of order 2.

3.3.5 The Isomorphism Theorems

The Isomorphism Theorems for groups look just like the versions for rings.

Theorem 3.24 (First Isomorphism Theorem) *Let G_1 and G_2 be groups, and let $\theta : G_1 \to G_2$ be a homomorphism. Then*

(a) $\mathrm{Im}(\theta)$ *is a subgroup of G_2;*

(b) $\mathrm{Ker}(\theta)$ *is a normal subgroup of G_1;*

(c) $G_1/\mathrm{Ker}(\theta) \cong \mathrm{Im}(\theta)$.

Proof We already proved the first two parts of this theorem. We have to prove (c). That is, we have to construct a bijection ϕ from G/N to $\mathrm{Im}(\theta)$, where $N = \mathrm{Ker}(\theta)$, and prove that it preserves the group operation.

The map ϕ is defined by $(Nx)\phi = x\theta$. We have

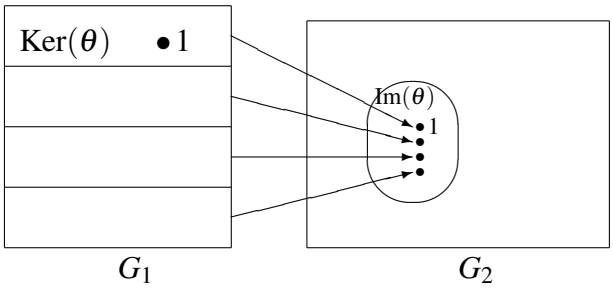
$$(Nx)\phi = (Ny)\phi \Leftrightarrow x\theta = y\theta \Leftrightarrow (xy^{-1})\theta = 1 \Leftrightarrow xy^{-1} \in \mathrm{Ker}(\theta) = N \Leftrightarrow Nx = Ny,$$

so ϕ is well-defined and one-to-one. It is clearly onto. Finally,

$$(Nx)\phi \cdot (Ny)\phi = (x\theta)(y\theta) = (xy)\theta = (Nxy)\phi = ((Nx)(Ny))\phi,$$

so ϕ preserves the group operation as required.

The same picture as for rings may be useful:



The parts on the left are the cosets of $N = \mathrm{Ker}(\theta)$, where N itself is the topmost part. Each coset of N maps to a single element of $\mathrm{Im}(\theta)$, and the correspondence between cosets and elements of $\mathrm{Im}(\theta)$ is the bijection of the last part of the theorem.

The other two theorems will be stated without proof. You are encouraged to try the proofs for yourself; they are very similar to the proofs for rings.

Theorem 3.25 (Second Isomorphism Theorem) *Let N be a normal subgroup of the group G. Then there is a one-to-one correspondence between the subgroups of G/N and the subgroups of G containing N, given as follows: to a subgroup H of G containing N corresponds the subgroup H/N of G/N. Under this correspondence, normal subgroups of G/N correspond to normal subgroups of G containing N; and, if M is a normal subgroup of G containing N, then*

$$(G/N)/(M/N) \cong G/M.$$

For the next theorem we need to define, for any two subsets A, B of a group G, the set

$$AB = \{ab : a \in A, b \in B\}$$

of all products of an element of A by an element of B.

Theorem 3.26 (Third Isomorphism Theorem) *Let G be a group, H a subgroup of G, and N a normal subgroup of G. Then*

(a) HN is a subgroup of G containing N;

(b) $H \cap N$ is an normal subgroup of H;

(c) $H/(H \cap N) \cong (HN)/N$.

We end this section with one fact about groups which doesn't have an obvious analogue for rings.

Proposition 3.27 *Let H and K be subgroups of the group G.*

(a) HK is a subgroup of G if and only if $HK = KH$.

(b) If K is a normal subgroup of G, then $HK = KH$.

Proof (a) Suppose that $HK = KH$. Then every element of the form kh (for $k \in K$ and $h \in H$) can also be expressed in the form $h'k'$ (for $h' \in H$ and $k' \in K$). Now we apply the subgroup test to HK. Take $h_1k_1, h_2k_2 \in HK$. We want to know if $h_1k_1(h_2k_2)^{-1} \in HK$. This expression is $h_1k_1k_2^{-1}h_2^{-1}$. Now $k_1k_2^{-1} \in K$, so $(k_1k_2^{-1})h_2^{-1} \in KH$; so we can write this element as $h'k'$, for some $h' \in H$ and $k' \in K$. Then

$$h_1k_1k_2^{-1}h_2^{-1} = (h_1h')k' \in HK,$$

as required.

Conversely, suppose that HK is a subgroup. We have to show that $HK = KH$, that is, every element of one is in the other. Take any element $x \in HK$. Then $x^{-1} \in HK$, so $x^{-1} = hk$, for some $h \in H$ and $k \in K$. Then $x = k^{-1}h^{-1} \in KH$. The reverse inclusion is proved similarly.

(b) If K is a normal subgroup, then the Third Isomorphism Theorem shows that HK is a subgroup, so that $HK = KH$ by Part (a).

Exercise If H and K are subgroups of G, show that

$$|HK| = \frac{|H| \cdot |K|}{|H \cap K|},$$

whether or not HK is a subgroup. [Hint: there are $|H| \cdot |K|$ choices of an expression hk. Show that every element in HK can be expressed as such a product in $|H \cap K$ different ways.]

Example Let $G = S_3$, $H = \{1, (1,2)\}$ and $K = \{1, (2,3)\}$. Then H and K are two subgroups of G, each of order 2, and $H \cap K = \{1\}$, so $|HK| = 4$. Since 4 doesn't divide 6, Lagrange's Theorem shows that HK cannot be a subgroup of G. This shows, once again, that H and K are not normal subgroups of G.

3.3.6 Conjugacy

Conjugacy is another equivalence relation on a group which is related to the idea of normal subgroups.

Let G be a group, we say that two elements $g, h \in G$ are *conjugate* if $h = x^{-1}gx$ for some element $x \in G$.

Proposition 3.28 *(a) Conjugacy is an equivalence relation on G.*

(b) A subgroup H of G is normal if and only if it is a union of (some of the) conjugacy classes in G.

Proof (a) Write $g \sim h$ to mean that $h = x^{-1}gx$ for some $x \in G$. Then

- $g = 1^{-1}g1$, so $g \sim g$: $\sim$ is reflexive.
- If $h = x^{-1}gx$, then $g = (x^{-1})^{-1}h(x^{-1})$: so $\sim$ is symmetric.
- Suppose that $g \sim h$ and $h \sim k$. Then $h = x^{-1}gx$ and $k = y^{-1}hy$ for some x, y. Then $k = y^{-1}x^{-1}gxy = (xy)^{-1}g(xy)$, so $g \sim k$: $\sim$ is transitive.

(b) The condition that H is a union of conjugacy classes means that, if $h \in H$, then any element conjugate to h is also in H. We saw in Proposition 3.19 that this condition is equivalent to normality of H.

Exercise Let $G = S_3$. Show that the conjugacy classes in G are $\{1\}$, $\{(1,2,3),(1,3,2)\}$, and $\{(1,2),(2,3),(1,3)\}$. (We will look at conjugacy in symmetric groups in the next section.)

3.4 Symmetric groups and Cayley's Theorem

Cayley's Theorem is one of the reasons why the symmetric groups form such an important class of groups: in a sense, if we understand the symmetric groups completely, then we understand all groups!

Theorem 3.29 *Every group is isomorphic to a subgroup of some symmetric group.*

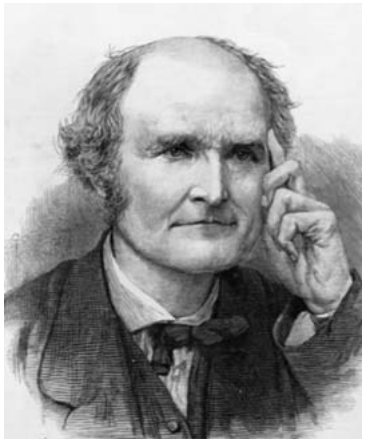
Before we give the proof, here is a small digression on the background. Group theory began in the late 18th century: Lagrange, for example, proved his theorem in 1770. Probably the first person to write down the axioms for a group in anything like the present form was Dyck in 1882. So what exactly were group theorists doing for a hundred years?

The answer is that Lagrange, Galois, etc. regarded a group as a set G of permutations with the properties

- G is closed under composition;
- G contains the identity permutation;
- G contains the inverse of each of its elements.

In other words, G is a subgroup of the symmetric group.

Thus, Cayley's contribution was to show that every group (in the modern sense) could be regarded as a group of permutations; that is, every structure which satisfies the group axioms can indeed be thought of as a group in the sense that Lagrange and others would have understood.
In general, systems of axioms in mathematics are usually not invented out of the blue, but are an attempt to capture some theory which already exists.



3.4.1 Proof of Cayley's Theorem

We begin with an example. Here is the Cayley table of a group we have met earlier: it is $C_2 \times C_2$, or the additive group of the field of four elements. When we saw it in Chapter 7, its elements were called e, a, b, c; now I will call them g_1, g_2, g_3, g_4.

$\cdot$	g_1	g_2	g_3	g_4
g_1	g_1	g_2	g_3	g_4
g_2	g_2	g_1	g_4	g_3
g_3	g_3	g_4	g_1	g_2
g_4	g_4	g_3	g_2	g_1

Now consider the four columns of this table. In each column, we see the four group elements $g_1, \ldots, g_4$, each occurring once; so their subscripts form a permutation of $\{1,2,3,4\}$. Let π_i be the permutation which is given by the ith column.

For example, for $i = 3$, the elements of the column are (g_3, g_4, g_1, g_2), and so π_3 is the permutation which is $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix}$ in two-line notation, or $(1,3)(2,4)$ in cycle notation.

The four permutations which arise in this way are:

$$\begin{aligned} \pi_1 &= 1 \\ \pi_2 &= (1,2)(3,4) \\ \pi_3 &= (1,3)(2,4) \\ \pi_4 &= (1,4)(2,3) \end{aligned}$$

Now we claim that $\{\pi_1, \pi_2, \pi_3, \pi_4\}$ is a subgroup H of the symmetric group S_4, and that the map θ defined by $g_i\theta = \pi_i$ is an isomorphism from G to H. (This means that, if $g_i g_j = g_k$, then $\pi_i \pi_j = \pi_k$, where permutations are composed in the usual way.) This can be verified with a small amount of checking.

You might think that it would be easier to use rows instead of columns in this argument. In the case of an Abelian group, like the one in this example, of course it makes no difference since the Cayley table is symmetric; but for non-Abelian groups, the statement would not be correct for rows.

So now we come to a more precise statement of Cayley's Theorem. We assume here that the group is finite; but the argument works just as well for infinite groups too.

Theorem 3.30 (Cayley's Theorem) *Let $G = \{g_1, \ldots, g_n\}$ be a finite group. For $j \in \{1, \ldots, n\}$, let π_j be the function from $\{1, \ldots, n\}$ to itself defined by the rule*

$$i\pi_j = k \text{ if and only if } g_i g_j = g_k.$$

Then

(a) π_j is a permutation of $\{1, \ldots, n\}$;

(b) the set $H = \{\pi_1, \ldots, \pi_n\}$ is a subgroup of S_n;

(c) the map $\theta : G \to S_n$ given by $g_j\theta = \pi_j$ is a homomorphism with kernel $\{1\}$ and image H;

(d) G is isomorphic to H.

Proof (a) To show that π_j is a permutation, it is enough to show that it is one-to-one, since a one-to-one function on a finite set is onto. (For infinite groups, we would also have to prove that it is onto.) So suppose that $i_1\pi_j = i_2\pi_j = k$. This means, by definition, that $g_{i_1}g_j = g_{i_2}g_j = g_k$. Then by the cancellation law, $g_{i_1} = g_{i_2} = g_k g_j^{-1}$, and so $i_1 = i_2$.

(c) Clearly the image of θ is H. So if we can show that θ is a homomorphism, the fact that H is a subgroup of S_n will follow.

Suppose that $g_j g_k = g_l$. We have to show that $\pi_j\pi_k = \pi_l$, in other words (since these are functions) that $(i\pi_j)\pi_k = i\pi_l$ for any $i \in \{1, \ldots, n\}$. Define numbers p, q, r by $g_i g_j = g_p$, $g_p g_k = g_q$, and $g_i g_l = g_r$. Then $i\pi_j = p$, $p\pi_k = q$ (so $i\pi_j\pi_k = q$), and $i\pi_l = r$. So we have to prove that $q = r$. But

$$g_q = g_p g_k = (g_i g_j)g_k = g_i(g_j g_k) = g_i g_l = g_r,$$

so $q = r$.

Now the kernel of θ is the set of elements g_j for which π_j is the identity permutation. Suppose that $g_j \in \mathrm{Ker}(\theta)$. Then $i\pi_j = i$ for any $i \in \{1, \ldots, n\}$. This means that $g_i g_j = g_i$, so (by the cancellation law) g_j is the identity. So $\mathrm{Ker}(\theta) = \{1\}$.

Now the First Isomorphism Theorem shows that $H = \mathrm{Im}(\theta)$ is a subgroup of S_n (that is, (b) holds), and that

$$G \cong G/\{1\} = G/\mathrm{Ker}(\theta) \cong H,$$

that is, (d) holds. So we are done.

Remark There may be other ways to find a subgroup of S_n isomorphic to the given group G. For example, consider the group $G = S_3$, a non-abelian group of order 6, whose Cayley table we wrote down in Chapter 7. From this table, you could find a set of six permutations in the symmetric group S_6 which form a subgroup of S_6 isomorphic to G. But G is already a subgroup of S_3!

3.4.2 Conjugacy in symmetric groups

We finish this chapter with two more topics involving symmetric groups. First, how do we tell whether two elements of S_n are conjugate?

We define the *cycle structure* of a permutation $g \in S_n$ to be the list of cycle lengths of g when it is expressed in cycle notation. (We include all cycles including those of length 1.) The order of the terms in the list is not important. Thus, for example, the permutation $(1,7)(2,6,5)(3,8,4)$ has cycle structure $[2,3,3]$.

Proposition 3.31 *Two permutations in S_n are conjugate in S_n if and only if they have the same cycle structure.*

Proof Suppose that $(a_1, a_2, \ldots, a_r)$ is a cycle of a permutation g. This means that g maps

$$a_1 \mapsto a_2 \mapsto \ldots \mapsto a_r \mapsto a_1.$$

We claim that $h = x^{-1}gx$ maps

$$a_1x \mapsto a_2x \mapsto \ldots \mapsto a_rx \mapsto a_1x,$$

so that it has a cycle $(a_1, a_2, \ldots, a_r)$. This is true because

$$(a_ix)h = (a_ix)(x^{-1}gx) = a_igx = a_{i+1}x$$

for $i = 1, \ldots, r-1$, and $(a_rx)h = a_1x$.

This shows that conjugate elements have the same cycle structure. The recipe is: given g in cycle notation, replace each element a_i in each cycle by its image under x to obtain $x^{-1}gx$ in cycle notation.

Conversely, suppose that g and h have the same cycle structure. We can write h under g so that cycles correspond. Then the permutation x which takes each point in a cycle of g to the corresponding point in a cycle of h is the one we require, satisfying $h = x^{-1}gx$, to show that g and h are conjugate.

Example The permutations $g = (1,7)(2,6,5)(3,8,4)$ and $h = (2,3)(1,5,4)(6,8,7)$ are conjugate. We can take x to be the permutation given by $\begin{pmatrix} 1 & 7 & 2 & 6 & 5 & 3 & 8 & 4 \\ 2 & 3 & 1 & 5 & 4 & 6 & 8 & 7 \end{pmatrix}$ in two-line notation, or $(1,2)(3,6,5,4,7)(8)$ in cycle notation.

3.4.3 The alternating groups

You have probably met the sign of a permutation in linear algebra; it is used in the formula for the determinant. It is important in group theory too.

Let g be an element of S_n, which has k cycles when written in cycle notation (including cycles of length 1). We define its *sign* to be $(-1)^{n-k}$.

Note that the sign depends only on the cycle structure; so *conjugate permutations have the same sign*.

Theorem 3.32 *The function* sgn *is a homomorphism from the symmetric group S_n to the multiplicative group $\{+1,-1\}$. For $n \geq 2$, it is onto; so its kernel (the set of permutations of $\{1,\ldots,n\}$ with sign $+$) is a normal subgroup of index 2 in S_n, called the* alternating group A_n.

Example For $n = 3$, the permutations with sign $+1$ are 1, $(1,2,3)$ and $(1,3,2)$, while those with sign -1 are $(1,2)$, $(2,3)$ and $(1,3)$. We have seen that the first three form a normal subgroup.

Proof We define a *transposition* to be a permutation of $\{1,\ldots,n\}$ which interchanges two points i and j and fixes all the rest. Now a transposition has cycle structure $[2,1,1,\ldots,1]$, and so has $n-1$ cycles; so its sign is $(-1)^1 = -1$.

We show the following two facts:

(a) Every permutation can be written as the product of transpositions.

(b) If t is a transposition and g any permutation, then $\mathrm{sgn}(gt) = -\mathrm{sgn}(g)$.

Now the homomorphism property follows. For take any $g,h \in S_n$. Write $h = t_1 t_2 \ldots t_r$, where $t_1,\ldots,t_r$ are transpositions. Then applying (b) r times, we see that $\mathrm{sgn}(gh) = \mathrm{sgn}(g)(-1)^r$. But also $\mathrm{sgn}(h) = (-1)^r$ (using the identity instead of g), so $\mathrm{sgn}(gh) = \mathrm{sgn}(g)\,\mathrm{sgn}(h)$. Thus sgn is a homomorphism. Since $\mathrm{sgn}(1,2) = -1$, we see that $\mathrm{Im}(\mathrm{sgn}) = \{+1,-1\}$. So, if A_n denotes $\mathrm{Ker}(\mathrm{sgn})$, the First Isomorphism Theorem shows that $S_n/A_n \cong \{\pm 1\}$, so that A_n has two cosets in S_n (that is, index 2).

Proof of (a): Take any permutation g, and write it in cycle notation, as a product of disjoint cycles. It is enough to show that each cycle can be written as a product of transpositions. Check that

$$(a_1,a_2,\ldots,a_r) = (a_1,a_2)(a_1,a_3)\cdots(a_1,a_r).$$

Proof of (b): Again, write g in cycle notation. Now check that, if t interchanges points in different cycles of g, then in the product gt these two cycles are "stitched together" into a single cycle; while, if t interchanges points in the same g-cycle, then this cycle splits into two in gt. For example,

$$\begin{aligned}(1,3,5,7)(2,4,6)\cdot(3,4) &= (1,4,6,2,3,5,7),\\ (1,2,5,3,8,6,4,7)\cdot(3,4) &= (1,2,5,4,7)(3,8,6).\end{aligned}$$

So multiplying g by a transposition changes the number of cycles by one (either increases or decreases), and so multiplies the sign by -1.

The proof shows another interesting fact about permutations. As we saw, every permutation can be written as a product of transpositions.

Corollary 3.33 *Given any two expressions of $g \in S_n$ as a product of transpositions, the numbers of transpositions used have the same parity, which is even if* $\mathrm{sgn}(g) = +1$ *and odd if* $\mathrm{sgn}(g) = -1$.

Proof We saw that if g is the product of r transpositions, then $\mathrm{sgn}(g) = (-1)^r$. This must be the same for any other expression for g as a product of transpositions.

Example $(1,2) = (1,3)(2,3)(1,3)$; one expression uses one transposition, the other uses three.

3.5 Some special groups

In the final section we make the acquaintance of some further types of groups, and investigate more closely the groups S_4 and S_5.

3.5.1 Normal subgroups of S_4 and S_5

In this section, we find all the normal subgroups of the groups S_4 and S_5. There are two possible approaches we could take. We could find all the subgroups and check which ones on our list are normal. But, for example, S_5 has 156 subgroups, so this would be quite a big job! The approach we will take is based on the fact that a subgroup of G is normal if and only if it is a union of conjugacy classes. So we find the conjugacy classes in each of these groups, and then figure out how to glue some of them together to form a subgroup (which will automatically be normal).

Recall from the last chapter that two permutations in S_n are conjugate if and only if they have the same cycle structure. So first we list the possible cycle structures and count the permutations with each structure. For S_4, we get the following table. (We list the sign in the last column; we know that all the permutations with sign $+1$ must form a normal subgroup. The sign is of course $(-1)^{n-k}$, where k is the number of cycles.)

Cycle structure	Number	Sign
$[1,1,1,1]$	1	$+$
$[2,1,1]$	6	$-$
$[2,2]$	3	$+$
$[3,1]$	8	$+$
$[4]$	6	$-$
Total	24	

How do we compute these numbers? There is a general formula for the number of permutations with given cycle structure. If you want to use it to check, here it is. Suppose that the cycle structure is $[a_1, a_2, \ldots, a_r]$, and suppose that in this list the number 1 occurs m_1 times, the number 2 occurs m_2 times, and so on. Then the number of permutations with this cycle structure is

$$\frac{n!}{1^{m_1} m_1! 2^{m_2} m_2! \cdots}.$$

So for example, for the cycle structure $[2,2]$, we have two 2s and nothing else, so $m_2 = 2$, and the number of permutations with cycle structure $[2,2]$ is $4!/(2^2 2!) = 3$.

In small cases we can argue directly. There is only one permutation with cycle structure $[1,1,1,1]$, namely the identity. Cycle structure $[2,1,1]$ describes transpositions, and there are six of these (the number of choices of the two points to be transposed). For cycle structure $[2,2]$ we observe that the six transpositions fall into three complementary pairs, so there are three such elements. For $[3,1]$, there are four choices of which point is fixed, and two choices of a 3-cycle on the remaining points. Finally, for $[4]$, a 4-cycle can start at any point, so we might as well assume that the first point is 1. Then there are $3! = 6$ ways to put the remaining points into a bracket $(1, \ , \ , \ )$ to make a cycle.

Having produced this table, how can we pick some of the conjugacy classes to form a subgroup? We know that a subgroup must contain the identity, so the first class must be included. Also, by Lagrange's Theorem, the order of any subgroup must divide the order of the group. So, unless we take all five classes, we cannot include a class of size 6. (For then the order of the subgroup would be at least 7, so necessarily 8 or 12, and we cannot make up 1 or 5 elements out of the remaining classes.) So the only possibilities are:

- Take just the class $\{1\}$. This gives the trivial subgroup, which is certainly normal.
- Take $\{1\}$ together with the class $[2,2]$, giving four elements. We have to look further at this.
- Take $\{1\}$ together with the classes $[2,2]$ and $[1,3]$. These are all the even permutations, so do form a normal subgroup, namely the alternating group A_4.
- Take all five classes. This gives the whole of S_4, which is a normal subgroup of itself.

The one case still in doubt is the possibility that the set

$$V_4 = \{1, (1,2)(3,4), (1,3)(2,4), (1,4)(2,3)\}$$

is a normal subgroup. Of course, if it is a subgroup, then it is normal, since it consists of two conjugacy classes. And it is a subgroup; our example of Cayley's Theorem produced precisely this subgroup! It is called V from the German word *vier*, meaning "four"; it is sometimes referred to as the "four-group".

We have proved:

Proposition 3.34 *The group S_4 has four normal subgroups. These are the identity, the four-group V_4, the alternating group A_4, and the symmetric group S_4.*

What about the factor groups? Clearly $S_4/\{1\} \cong S_4$, while $S_4/S_4 \cong \{1\}$. We know that S_4/A_4 is isomorphic to the multiplicative group $\{\pm 1\}$, which is a cyclic group of order 2. One case remains:

Proposition 3.35 $S_4/V_4 \cong S_3$.

Proof There are many ways to see this. Here is the simplest.

Consider the subgroup S_3 of S_4 consisting of all permutations fixing the point 4. We have $|S_3| = 6$, $|V_4| = 4$, and $S_3 \cap V_4 = \{1\}$ (by inspection of the elements of V_4), so $|S_3V_4| = 24$; that is, $S_3V_4 = S_4$. Now, by the Third Isomorphism Theorem,

$$S_4/V_4 = S_3V_4/V_4 \cong S_3/(S_3 \cap V_4) = S_3/\{1\} = S_3.$$

We now look at S_5 and show:

Proposition 3.36 *The group S_5 has three normal subgroups: the identity, the alternating group A_5, and the symmetric group S_5.*

We will be more brief here. The table of conjugacy classes looks like this:

Cycle structure	Number	Sign
$[1,1,1,1,1]$	1	$+$
$[2,1,1,1]$	10	$-$
$[2,2,1]$	15	$+$
$[3,1,1]$	20	$+$
$[3,2]$	20	$-$
$[4,1]$	30	$-$
$[5]$	24	$+$
Total	120	

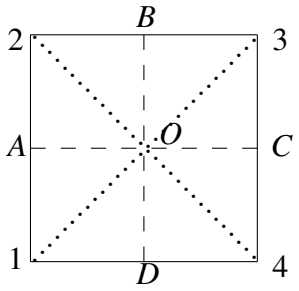
Number these classes as $C_1,\ldots,C_7$ in order. We have to choose some of them including C_1 such that the sum of the numbers is a divisor of 120. All classes except C_1 and C_7 have size divisible by 5; so if we don't include C_7 then the total divides 24, which is easily seen to be impossible. So we must have C_7. Now, since we are trying to build a subgroup, it must be closed under composition; so any cycle type which can be obtained by multiplying together two 5-cycles has to be included. Since

$$\begin{aligned}(1,2,3,4,5)(1,5,4,2,3) &= (1,3,2),\\ (1,2,3,4,5)(1,2,3,5,4) &= (1,3)(2,5),\end{aligned}$$

both classes C_3 and C_4 must be included. Now $C_1\cup C_3\cup C_4\cup C_7$ is the alternating group A_5, and if there is anything else, we must have the entire symmetric group.

3.5.2 Dihedral groups

One important source of groups is as symmetries of geometric figures. Here is an example. Consider a square, as shown in the figure. (We have marked various axes of symmetry as dotted lines.)



Now the square has eight symmetries, four rotations and four reflections. They are given in the table, together with their effect as permutations of the four vertices of the square. The rotations are taken clockwise.

Symmetry	Permutation
Identity	1
Rotation through 90° about O	$(1,2,3,4)$
Rotation through 180° about O	$(1,3)(2,4)$
Rotation through 270° about O	$(1,4,3,2)$
Reflection about AC	$(1,2)(3,4)$
Reflection about BD	$(1,4)(2,3)$
Reflection about 13	$(2,4)$
Reflection about 24	$(1,3)$

The eight symmetries form a group (with the operation of composition). The corresponding permutations form a group, a subgroup of the symmetric group S_4, which is isomorphic to the group of symmetries. This group is non-Abelian. (This can be seen by composing symmetries, or by composing permutations. For example, $(1,2,3,4)(1,3) = (1,2)(3,4)$, while $(1,3)(1,2,3,4) = (1,4)(2,3)$.)

More generally, a regular n-gon has $2n$ symmetries, n rotations and n reflections, forming a group which is known as the *dihedral group* D_{2n}. (Thus the group in the table above is D_8. You should be warned that some people refer to what I have called D_{2n} as simply D_n.)

Here are some properties of dihedral groups, which should be clear from the figure.

- The n rotations form a cyclic subgroup C_n of D_{2n}. (This subgroup has index 2 in D_{2n}, so it is a normal subgroup.)
- If n is odd, then every reflection axis joins a vertex to the midpoint of the opposite side; while if n is even, then $n/2$ axes join the midpoints of opposite sides and $n/2$ join opposite vertices.
- Any reflection has order 2.
- If a is a rotation and b a reflection, then $bab = a^{-1}$.

The last condition says: reflect the figure, rotate it clockwise, and then reflect again; the result is an anticlockwise rotation. This gives another proof that the rotations form a normal subgroup. For let a be a rotation, and x any element. If x is a rotation, then $ax = xa$, so $x^{-1}ax = a$. If x is a reflection, then $x^{-1} = x$, and so $x^{-1}ax = a^{-1}$. So any conjugate of a rotation is a rotation.

The definition of D_4 is not clear, since there is no regular 2-gon. But if we take the pattern in D_{2n} and apply it for $n = 2$, we would expect a group with a cyclic subgroup of order 2, and all elements outside this subgroup having order 2. This is a description of the four-group.

Moreover, D_6 is the group of symmetries of an equilateral triangle, and the six permutations of the vertices comprise all possible permutations.

So the following is true.

Proposition 3.37 *(a) The dihedral group D_4 is isomorphic to the four-group V_4.*

(b) The dihedral group D_6 is isomorphic to the symmetric group S_3.

3.5.3 Small groups

How many different groups of order n are there? (Here, "different" means "non-isomorphic".) This is a hard question, and the answer is not known in general: it was only six years ago that the number of groups of order 1024 was computed: the number is $49,487,365,422$. (The result of this computation was announced at Queen Mary.)

We will not be so ambitious. For small n, the number of groups of order n is given in this table. We will verify the table up to $n = 7$.

Order	1	2	3	4	5	6	7	8
Number of groups	1	1	1	2	1	2	1	5

Clearly there is only one group of order 1. The result for $n = 2,3,5,7$ follows from the next Proposition.

Proposition 3.38 *A group of prime order p is isomorphic to the cyclic group C_p.*

Proof Take an element g of such a group G, other than the identity. By Lagrange's Theorem, the order of g must divide p, and so the order must be 1 or p (since p is prime). But $g \neq 1$, so its order is not 1; thus it is p. So G is a cyclic group generated by g.

Next we show that there are just two groups of order 4. We know two such groups already: the cyclic group C_4, and the four-group V_4. (See the Cayley tables in 3.1.2 of the notes).

Let G be a group of order 4. Then the order of any element of G divides 4, and so is 1, 2 or 4 by Lagrange's Theorem. If there is an element of order 4, then G is cyclic; so suppose not. Then we can take $G = \{1,a,b,c\}$, where $a^2 = b^2 = c^2 = 1$. What is ab? It cannot be 1, since $ab = 1 = a^2$ would imply $a = b$ by the Cancellation Law. Similarly $ab \neq a$ and $ab \neq b$, also by the Cancellation Law (these would imply $b = 1$ or $a = 1$ respectively). So $ab = c$. Similarly, all the

other products are determined: the product of any two non-identity elements is the third. In other words, the Cayley table is

$\cdot$	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

We recognise the Klein four-group. So there is just one such group (up to isomorphism), giving two altogether.

To deal with order 6, we need a couple of preliminary results.

Proposition 3.39 *A group of even order must contain an element of order* 2*.*

Proof Take any group G of order n, and attempt to pair up the elements of G with their inverses. Suppose that we can form m pairs, accounting for $2m$ elements. The elements we have failed to pair up are the ones which satisfy $g = g^{-1}$, or $g^2 = 1$; these include the identity (one element) and all the elements of order 2. So there must be $n-2m-1$ elements of order 2. If n is even, then $n-2m-1$ is odd, and so cannot be zero; so there is an element of order 2 in G.

Proposition 3.40 *A finite group in which every element has order* 2 *(apart from the identity) is Abelian.*

Proof Take any $g,h \in G$. We have

$$\begin{aligned} (gh)^2 = ghgh &= 1, \\ g^2h^2 = gghh &= 1, \end{aligned}$$

so by cancellation, $hg = gh$. Thus G is Abelian.

Now let G be a group of order 6. If G contains an element of order 6, then it is cyclic; so suppose not. Now all its elements except the identity have order 2 or 3. The first proposition above shows that G contains an element a of order 2. The second shows that it must also have an element of order 3. For, suppose not. Then all non-identity elements of G have order 2. If g and g are two such elements, then it is easy to see that $\{1,g,h,gh = hg\}$ is a subgroup of order 4, contradicting Lagrange's Theorem.

So let a be an element of order 3 and an element b of order 2. The cyclic subgroup $\langle a\rangle = \{1,a,a^{-1} = a^2\}$ of G has order 3 and index 2, so it is normal. So $b^{-1}ab \in \langle a\rangle$, whence $b^{-1}ab = a$ or $b^{-1}ab = a^{-1}$.

If $b^{-1}ab = a$, then $ab = ba$, so $(ab)^i = a^i b^i$ for all i. Then the powers of ab are

$$(ab)^2 = a^2b^2 = a^2, \quad (ab)^3 = a^3b^3 = b, \quad (ab)^4 = a^4b^4 = a,$$
$$(ab)^5 = a^5b^5 = a^2b, \quad (ab)^6 = a^6b^6 = 1,$$

so the order of ab is 6, contradicting our case assumption. So we must have $b^{-1}ab = a^{-1}$, or $ba = a^{-1}b = a^2b$.

Now using this, the Cayley table of G is completely determined: all the elements have the form a^ib^j, where $i = 0,1,2$ and $j = 0,1$; to multiply a^ib^j by a^kb^l, we use the condition $ba = a^{-1}b$ to jump the first b over the as to its right if necessary and the conditions $a^3 = b^2 = 1$ to reduce the exponents. For example,

$$a^2b \cdot ab = a^2(ba)b = a^2(a^2b)b = a^4b^2 = a.$$

So there is only one possible group of this type. Its Cayley table is:

$\cdot$	1	a	a^2	b	ab	a^2b
1	1	a	a^2	b	ab	a^2b
a	a	a^2	1	ab	a^2b	b
a^2	a^2	1	a	a^2b	b	ab
b	b	a^2b	ab	1	a^2	a
ab	ab	b	a^2b	a	1	a^2
a^2b	a^2b	ab	b	a^2	a	1

These relations are satisfied in the group S_3, if we take $a = (1,2,3)$ and $b = (1,2)$. So there is such a group; and so there are two groups of order 6 altogether (the other being the cyclic group). Alternatively, we could observe that the above relations characterise the dihedral group D_6, so the two groups are C_6 and D_6.

We see, incidentally, that $S_3 \cong D_6$, and that this is the smallest non-Abelian group.

3.5.4 Polyhedral groups

You have probably seen models of the five famous regular polyhedra: the tetrahedron, the cube (or hexahedron), the octahedron, the dodecahedron, and the icosahedron. These beautiful figures have been known since antiquity. See Figures 3.1, 3.2.

What are their symmetry groups?

Here I will just consider the groups of rotations; the extra symmetries realised by reflections in three-dimensional space make the situation a bit more complicated. As in the case of the dihedral groups, these groups can be realised as

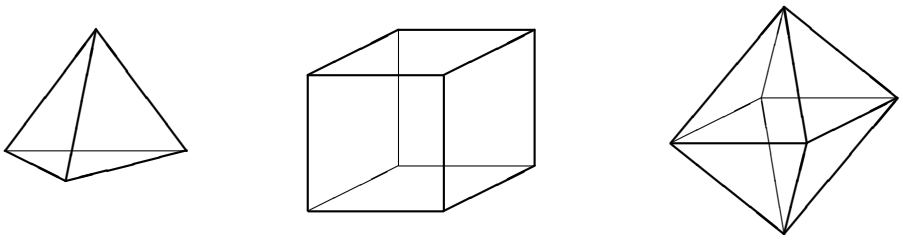


Figure 3.1: Tetrahedron, cube, and octahedron

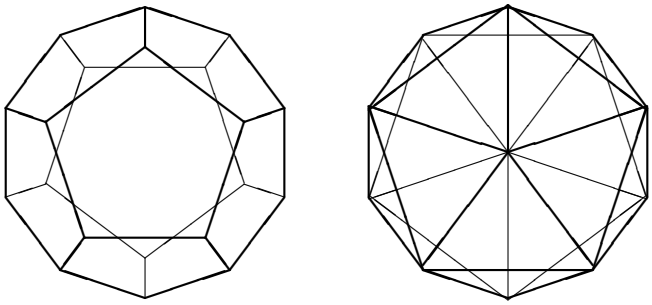


Figure 3.2: Dodecahedron and Icosahedron

permutation groups, by numbering the vertices and reading off the permutation induced by any symmetry.

Moreover, there are five figures, but only three groups. Apart from the tetrahedron, the figures fall into "dual pairs": the figure whose vertices are the face centres of a cube is an octahedron and *vice versa*, and a similar relation holds between the dodecahedron and the icosahedron. Dual pairs have the same symmetry group. (The face centres of the tetrahedron are the vertices of another tetrahedron, so this figure is "self-dual".) The following result describes the three symmetry groups.

Proposition 3.41 *(a) The tetrahedral group is isomorphic to A_4.*

(b) The octahedral group is isomorphic to S_4.

(c) The icosahedral group is isomorphic to A_5.

Proof I will outline the proof. First we compute the orders of these groups. If a figure has m faces, each a regular polygon with n sides, then the number of rotational symmetries is mn. For imagine the figure with one face on the table. I

can pick it up and rotate it so that any of the m faces is on the table, in any of the n possible orientations. Now we have

- Tetrahedron: $m = 4$, $n = 3$, group order 12.
- Cube: $m = 6$, $n = 4$, group order 24.
- Octahedron: $m = 8$, $n = 3$, group order 24.
- Dodecahedron: $m = 12$, $n = 5$, group order 60.
- Icosahedron: $m = 20$, $n = 3$, group order 60.

We see that the symmetry groups of dual polyhedra have the same order, as they should.

(a) Any symmetry of the tetrahedron permutes the four vertices. So the symmetry group is a subgroup of S_4 of order 12. To see that it is A_4, we simply have to observe that every symmetry is an even permutation of the vertices. A rotation about the line joining a vertex to the midpoint of the opposite face has cycle structure $[3,1]$, while a rotation about the line joining the midpoints of opposite edges has cycle structure $[2,2]$. (Alternatively, a subgroup of S_4 of order 12 has index 2 and is therefore normal; and we have worked out the normal subgroups of S_4. The only one of order 12 is A_4.)

(b) Consider the cube. It has four diagonals joining opposite vertices. Any symmetry induces a permutation of the four diagonals. It is not hard to see that the map from symmetries to permutations is one-to-one. So the group is isomorphic to a subgroup of S_4 of order 24, necessarily the whole of S_4.

(c) This is the hardest to prove. But in fact it is possible to embed a cube so that its vertices are eight of the 20 vertices of the dodecahedron in five different ways. These five inscribed cubes are permuted by any rotation, so we have a subgroup of S_5 of order 60. This subgroup has index 2 and so is normal; so it must be A_5.

Index